STATISTICAL METHODS AND ANALYSIS FOR GENOME-WIDE ASSOCIATION STUDIES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Lin Li

May 2010

STATISTICAL METHODS AND ANALYSIS FOR GENOME-WIDE ASSOCIATION STUDIES

Lin Li, Ph.D.

Cornell University 2010

Genome-wide association (GWA) studies utilize a large number of genetic variants, usually single nucleotide polymorphisms (SNPs), across the entire genome to identify genetic basis underlying disease susceptibility or phenotypic variation in a trait of interest. A commonly used analysis tool is single marker analysis (SMA), which tests one SNP at a time. Although it has been successful in identifying some causal loci, further enhancements are possible by considering multi-locus methods that investigate a large number of SNPs simultaneously. One difficulty of doing so is high dimensionality, i.e. the large number of SNPs, making it a challenging statistical problem. My first project addresses this problem in case-control GWA studies. Both the logistic and probit models are considered for binary traits, and three-component mixture priors are assumed to model the fact that only a few SNPs have non-negligible effects. To estimate posterior distributions, I propose three Markov chain Monte Carlo techniques. Specifically, an adaptive independence sampler is proposed for the logistic model, and data augmentation methods are developed for both logistic and probit models. Simulations suggest that they nearly always outperform SMA. The second project deals with GWA studies on quantitative traits with the confounding of population structure. A linear mixed model is used to account for cryptic relatedness between individuals in the sample. I propose an algorithm that is based on least angle regression and can efficiently select a

small number of SNPs that are likely to be associated with the trait. Simulations show that the proposed algorithm tends to yield higher ranks for causal loci than least angle regression directly applied, and that both outperform SMA. My third project is part of the so-called CanMap project. More than 1,000 domestic dogs from different breeds, wild canids and village dogs were genotyped on a dense SNP array, and my responsibility was to carry out a GWA analysis for the domestic dog on body weight and other morphological traits including height, shapes, etc. The GWA results enrich our understanding of the impact of strong directional selection on the genetic architecture of complex traits known to be under selection.

BIOGRAPHICAL SKETCH

Lin Li began his undergraduate studies in the School of Mathematical Sciences at Peking University, Beijing, China. After earning a Bachelor's degree in Statistics, he continued in the same school to study in a master's program under the supervision of Professor Xiangzhong Fang. In 2005, right after obtaining his Master's degree in Statistics, Lin began studying in the Ph.D. program in Biostatistics at Emory University. With a great interest in genomics, he transferred to the Ph.D. program in Computational Biology at Cornell University in 2006. Ever since then he has been studying and doing research under the supervision of Professor Carlos D. Bustamante. After spending four years at Cornell University, Lin expects to continue his research as a postdoctoral research fellow at Harvard University Department of Biostatistics.

To my family.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the support and encouragement from faculty, colleagues and family members. I would like to take this precious opportunity to express my gratitude to them. Although with limited space I can not name everyone here, my thanks go to all of them and it has been great to have them on my road to a science Ph.D.

My sincere gratitude goes to my Ph.D. advisor, Professor Carlos Bustamante. Carlos has guided my research throughout my studies at Cornell, not only by discussing general ideas with me and giving constructive suggestions, but also by helping me tackle concrete problems. His kindness, patience and humor have made my research a very enjoyable experience. I am often inspired by his enthusiasm and energy in science, which set a successful example for me to pursue my future career. Besides advising my research, he has also been very supportive to me and my family, and is greatly appreciated by us.

My committee members have played an important role for my Ph.D. training. I would like to thank Professor Andrew Clark. I still clearly remember the email of admission he sent to me that marked the start of my journey at Cornell. Ever since then, he has always been ready for offering help and thoughtful insights. For example he worked at weekend so that comments on my manuscript could go back to me promptly. I would also like to thank Professor Martin Wells, for sharing his expertise in statistics. Through valuable discussions, I have benefited a lot from his insights and sharpness, and his comments have often triggered new ideas in developing methodology. I would also like to thank Professor Ping Li. His expertise in computer science has let me better understand machine learning techniques. With his help, I was able to enrich my experience in reviewing papers and writing grant proposals.

Many other faculty members have also been very helpful and supportive. I would like to thank Professor Jason Mezey, for sharing his expertise in quantitative genomics. The lab meetings and many discussions with him have taught me a lot. He has also been very kind by offering me many considerate arrangements. I would also like to thank Professor Adam Siepel. His interesting, motivating and well-designed course on computational genomics laid a foundation for my following projects and will be indispensable for any future research.

I would also like to thank my master's advisor, Professor Xiangzhong Fang, at Peking University. It was he who directed me in my first-ever project in statistical genetics. He has been always supportive not only in the master's program but also during my graduate studies at Emory and Cornell.

My thanks also go to my colleagues and friends–to name a few–Keyan Zhao, Hong Gao, Chuan Gao, Abra Brisbin, Adam Boyko, Jeremiah Degenhardt, Benjamin Logsdon, Amit Indap, Fangfei Ye, Carly Hills, Beatrix Johnson, and many other kindly and supportive people in the department.

I would also like to express my gratitude to several parties that are indispensable to my research. The work was supported by NIH grant R01 HL084706 to Andrew Clark, Carlos Bustamante, Rasmus Nielsen, Simon Tavere, and Manolis Dermitzakis. Many simulations in my projects were run on the Apple cluster in the Department of Molecular Biology and Genetics as well as the BioSim cluster in the Department of Computer Science, whose technology support teams have offered me many helps.

Last but not the least, I would like to thank all my family members for their support. Especially, I am grateful to my wife, Jie, our parents, our son, Daniel, for their care and love. Without them, it would be impossible for me to complete this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

At the end of the twentieth century, people still did not have much knowledge of their genomes and could hardly imagine the accomplishments achieved in the following ten years. The Human Genome Project, drafted in 2001 [4, 5] and completed in 2003 [6], was the first international effort to sequence the human genome. With the aid of fast developing sequencing, genotyping and other biotechnology, researchers were able to decode genomes both more rapidly and at a larger scale. The International HapMap Project [7, 8, 9], with phase II completed in 2007, developed a haplotype map of the human genome by investigating hundreds of individuals in different populations. Launched in 2008, the 1000 Genomes Project aims to provide a more in-depth resource on human genetic variation [10] by sequencing more than 1,000 individuals. Besides humans, many other species have their genomes recently sequenced, including rice [11, 12], chickens [13], dogs [14], chimpanzees [15], macaques [16], cattle [17], etc. All the accomplishments have enriched people's knowledge of these species and provided valuable resources for scientific research.

1.1 Genetic basis of complex traits and genome-wide association studies

One important and active research area is to understand the genetic basis of complex traits and particularly the inherited causes of common diseases. The common disease, common variant hypothesis (CD/CV) stated that common diseases may be caused by a few common allelic variants [18]. Based on this

theory, the idea of genome-wide association (GWA) studies was proposed and discussed in the late 1990's [19, 20, 21]. As defined by the National Institutes of Health (http://grants.nih.gov/grants/gwas/), a GWA study is any study of genetic variation across the entire human genome that is designed to identify genetic associations with observations, or the presence or absence of a disease or condition. The single nucleotide polymorphism (SNP) is the most common form of genetic variation, millions of which scatter across the human genome. A mutation is more likely to travel with SNPs in close neighborhood than those residing farther away when passed from parent to offspring, which is called linkage disequilibrium (LD). If a mutation becomes relatively common in a population but remains in its LD block, any SNP within the same block may help identify the causative mutation [22]. With hundreds of thousands of SNPs genotyped and tested for associations across the whole genome, causal variants are hoped to be identified through the SNPs associated with the trait of interest. The trait can be a complex, non-Mendelian disease (e.g. type 1 diabetes, type 2 diabetes, breast cancer, etc.), or any measurable trait (e.g. height, lipids, fat mass, etc.).

The advance of cost-efficient high throughput genotyping technology and the availability of many reference resources have made GWA studies possible, and many such studies have been underway in the past few years. Early success of understanding diseases to which there is a genetic predisposition included several studies reported in 2005, where researchers identified causal variants for age-related macular degeneration [23, 24, 25]. In 2007, a milestone for GWA studies came from the Wellcome Trust Case-Control Consortium which identified dozens of genetic variants contributing to susceptibility of seven diseases [26]. As a summary, the National Human Genome Research Institute (NHGRI)

maintains an online catalog of published GWA studies for human diseases and traits (http://www.genome.gov/26525384). There are also GWA studies on other species such as dogs [3] and Arabidopsis thaliana [27]. All these studies have accelerated the search for genetic contributions to complex traits, especially those of humans.

It should also be noted that GWA studies may be ineffective if the disease is caused by some rare mutations [22, 28]. GWA studies rely on the CD/CV hypothesis; different from CD/CV, the rare variant (CD/RV) hypothesis states that rare variants account for the genetic susceptibility of diseases. While neither hypothesis always holds true, it has been proposed to integrate both hypotheses and construct a composite encompassing all influential genes for a multifactorial trait [18]. Given a trait, the contributions of common and rare causal variants will vary, and current GWA studies are able to identify the common ones while lacking of enough power to identify rare ones.

More detailed reviews on GWA studies can be found in recent literature [29, 30, 31, 32].

1.2 Population stratification and cryptic relatedness

A potential major problem in GWA studies is population structure, which can include stratification (individuals in the sample coming from different populations), and cryptic relatedness (unknown genetic relationship between individuals). The confounding, if not carefully accounted for, can result in spurious associations [33, 34] and hence elevate false positive rates [19, 35].

To account for population stratification and cryptic relatedness, various statistical methods have been proposed including genomic control (GC) [36], structured association (SA) [37], and principal component analysis (PCA) [38], etc. Recently, a unified linear mixed model approach (LMM) has been proposed[39], which can model stratification in fixed effects and model relatedness in random effects whose covariance matrix is some similarity matrix. LMM appears very promising in that it can account for stratification and relatedness simultaneously, and also in that it is able to take advantages of SA and PCA. A lot of interests have been drawn to LMM in the past few years, with research investigating its effectiveness [39, 40], the choice of the covariance matrix [39, 40, 41], and computational efficiency [41, 42, 43].

The problem of heavy computational burden is especially pressing for LMM in the context of GWA studies. LMM is more computationally intensive than linear models, and the difference can be dramatic if thousands of individuals and each with hundreds of thousands of SNPs, often found in GWA studies, are to be studied. Computational efficiency, therefore, becomes a top priority and should be kept in mind for researchers should they want to develop methods based on LMM.

1.3 Statistical challenges of high dimensionality

By definition, GWA studies can have hundreds of thousands of SNPs involved, and such a high dimension was hardly seen in classical statistics research and applications. The dimension is so haunting that researchers usually rely on single marker analysis that tests one SNP at a time (e.g. simple regression), even

though the ideal analysis should consider all or many SNP genotypes simultaneously. This brings challenges as well as opportunities for statisticians to develop advanced methods for high dimensional data. Such methods would need to consider multiple loci simultaneously, and also reduce the high dimension to a reasonable size. Many solutions have been proposed, including penalized likelihood methods [44], regularized estimating equations[45], Bayesian shrinkage analysis [46] and Bayesian variable selection [47], and so on. These methods, if applied to GWA data sets, could possibly improve performance over single-SNP tests, as weaker effects now condition on other effects and may appear more apparent [48]. Actually, some applications have been proved to be successful for genetic studies [49, 48, 50, 51].

Although active research on dimension reduction is underway in the statistical community and many methods have been introduced and applied to GWA studies, a great need of multi-locus methods and variable selection exists especially for binary traits and when accounting for population structure in GWA studies. Rather than ordinary linear models, generalized linear models and linear mixed models are usually used for these cases. It is both necessary and challenging to consider variable selection for these models, as generalized linear models lack of many computational advantages found in linear models, and linear mixed models are complicated with random effects, demanding highly computation-efficient methods.

1.4 Morphological traits in domestic dogs

The dog was domesticated from the wolf, which may date back to 15,000 years ago or even more, depending on the locations [14]. The "man's best friend" has a long history of being subjected to artificial selection from its companion, resulting in the formation of more than 300 dog breeds throughout the world. Having been selected mostly on morphology and behavior, the domestic dog can vary dramatically in its body size and shape and other traits among different breeds. Meanwhile, the power of humans reshaping the dog genome is so strong that the dog has formed significant inter-breed heterogeneity and intra-breed homogeneity regarding its genetic diversity [52]. These observations have drawn the interest of scientists to decode the genetic basis of morphological diversity in the domestic dog.

The characteristics of the dog genome and its phenotypic variation suggest that GWA studies can be done with as few as 20,000-30,000 SNPs and a small number of samples are required [14, 53]. The investment of such studies would be much less than that required for a human study, and yet still yields important results. The results can provide valuable guidance to dog owners and breeders, and be of great interest to a large community, given the dog's popularity in many cultures. The dog GWA studies can also have important implications for human health. The dog genome can be compared to the human genome, and studying genes responsible for traits in dogs can provide a valuable approach for better understanding human genetics.

To make all the GWA studies possible and successful, the dog genome was sequenced in 2005 [14], and another great effort, termed the CanMap Project, has

just been completed. The CanMap project has generated a data set consisting of dense SNP profiles of a dozen dogs from each of around 80 breeds. The data set is expected to be a resource of a great value, and can be used in many dog GWA studies.

1.5 Outline of the dissertation

The rest of this dissertations covers three projects I have been involved on statistical methods and applications for GWA studies.

Chapter 2 describes a Bayesian framework, termed Bayesian mixture models, to address the problem of high dimensionality in GWA studies. Emphasis is put on the case-control study design, popular in epidemiology and GWA studies, where a binary trait, e.g. the presence or absence of a condition or disease, is often involved. Generalized linear models, including the logistic and probit models, are used with three-component mixture priors assumed for marker effects to model the fact that only a small number of markers (usually SNPs) have non-negligible effects. Posterior distributions are estimated using three sets of Markov chain Monte Carlo techniques. Specifically, an adaptive independence sampler is proposed for the logistic model, and data augmentation introducing liabilities and a threshold is developed for both logistic and probit models. Simulations suggest that the proposed methods nearly always outperform single marker analysis which tests one marker at a time for associations.

Chapter 3 introduces an efficient linear mixed model that accounts for population stratification and genetic relatedness in association studies. A random effect whose correlation matrix is set to be a similarity matrix between individ-

uals is used to model the genetic relatedness. After variance parameters are estimated for a mixed model with no marker included, the residuals of the phenotypes are calculated. Then a least angle regression is run on the residuals as well as the genotypes searching for a parsimonious model of associations. As a comparison, a naive algorithm directly applies the least angle regression to the original phenotypes and genotypes. Simulations show that the proposed algorithm yields higher ranks for causal trait loci than the naive one, while both algorithms outperform single marker analysis that overlooks genetic relatedness. The proposed method was also applied to a dog study.

Chapter 4 discusses a statistical analysis for which I was responsible and that has been carried out in the CanMap Project. The project genotyped more than 1,000 dogs from different breeds on a dense SNP array. Genome-wide association studies were carried out for body weight and other morphological traits including height, shapes, etc. Detailed statistical methods and analysis results are reported in the chapter. Some of the methods discussed in Chapter 2 and 3 are also revisited as applications.

CHAPTER 2

BAYESIAN MIXTURE MODELS FOR CASE-CONTROL GENOME-WIDE ASSOCIATION STUDIES

2.1 Introduction

Genome-wide association (GWA) studies have the capacity to identify common genetic variants with modest effects on disease susceptibility [19, 20, 21]. Most GWA studies use the case-control experimental design [30], where a group of affected individuals (cases) and a group of unaffected ones (control) are compared to identify disease susceptibility alleles. For example, the Wellcome Trust Case-control Consortium [26] used 4,000 cases of seven common diseases and 3,000 shared controls to identify dozens of alleles influencing common chronic diseases. While these experiments have demonstrated the general feasibility of GWA mapping, there remains room for improvement in the analytical approaches used to identify putative marker-phenotype associations.

Most GWA studies still rely on single marker analysis (SMA) where each genotyped SNP is individually queried for association with disease outcome. This approach has gained wide acceptance because of its computational simplicity. However, SMA only investigates marginal distributions of SNPs and uses the information to prioritize these SNPs, while there are usually many genes underlying complex traits. SMA seems less informative than multi-SNP methods in the sense that weaker SNP effects would be more apparent after other effects have already been accounted for [48]. Secondly, another challenge for GWA is the so-called "curse of dimensionality" where the number of parameters to be estimated p (i.e. association metrics at millions of SNPs) is much larger than

the number of observations n (i.e. number of individuals typed at the markers). This leads to poor performance of classical likelihood based estimators of association and makes model selection (i.e., consideration of multi-SNP models of disease) a computationally challenging problem. Thirdly, multiple testing remains elusive in the GWA context since tests are correlated among markers both due to physical linkage among SNPs and the aforementioned "curse of dimensionality". While all three of these problems appear on the surface to be unrelated, they are, in fact, closely linked to the central issue of constructing an efficient predictive model of association between genotype and phenotype based on a matrix of genotypes and a vector of phenotypes (or even a matrix of phenotypes). At its core, this is a so-called dimension reduction problem where one wishes to reduce (or filter) the number of explanatory variables or features (in our case, SNPs) needed to predict a given outcome variable (in our case, disease status).

In the statistical community, there have been many efforts on tackling the problem of dimension reduction. One classical method is stepwise regression, which relies on some information criteria such as Akaike information criterion (AIC) [54] and Bayesian information criterion (BIC) [55]. Modern dimension reduction techniques can be loosely classified into three categories: 1) penalized likelihood methods, which is to add a penalty to the likelihood for modeling the sparseness. Examples include least absolute shrinkage and selection operator (LASSO) [44], adaptive LASSO [56], LASSO penalized logistic regression [50], and smoothly clipped absolute deviation [57]; 2) regularized estimating equation methods, like the Dantzig selector [45, 58]; 3) Bayesian methods, which use special priors for the explanatory variables to model the sparseness. Some early examples of such priors include the "spike and slab" prior [59] and stochastic

search variable selection (SSVS) [47]. Methods in these three categories have connections with each other [60, 58], for example, some penalty on likelihood can be regarded as some prior and the penalized likelihood can be connected to a posterior distribution function.

Some of these methods are deterministic, that is, always producing the same output given a particular input. Stepwise regression, categories 1), 2), and some Bayesian methods in 3) are all deterministic algorithms. Some methods have been applied to GWA studies that search for either likelihood maxima [50] or posterior modes[48, 51]. While some of them can be slow (e.g. stepwise) and some can be relatively fast (e.g. LASSO), a key problem with deterministic algorithms is that they tend to be stuck in locally optimal models. Some methods have to be applied several times with different initial values or use perturbations [48] hoping to make the problem less severe.

As opposed to deterministic, some Bayesian methods are stochastic. These methods usually rely on Markov chain Monte Carlo techniques to simulate the posterior distributions, and are in theory more likely to find the global optima. Although computationally intensive, these methods have another advantage of making the posterior distribution available, which can be used to assess the uncertainty of models. Given the fast increase of computing power nowadays, one may want to invest time and resource for the statistical inference. Many of such methods have already been applied to genetic studies. While special priors are usually assumed to model the "sparseness" fact that only a small number of loci are expected to have non-negligible effects on the trait, some methods stochastically search variables to be included in the model (i.e. variable selection), while others shrink effect sizes of markers that have negligible effect

on the trait towards zero (i.e shrinkage method). For example, Yi et al. [61, 62] and Zhang et al. [63, 64] use variable selection in linear models to identify quantitative trait loci; Huang et al. [49] apply the shrinkage method [46] to map multiple QTLs of complex binary traits in experimental crosses.

Given the current research activities, there are still tremendous interests in investigating Bayesian variable selection methods for binary traits, as usually found in case-control studies, in the context of GWA studies. The following sections of this chapter is to describe a Bayesian framework termed Bayesian mixture model focusing on generalized linear models for binary traits. More specifically, two popular models, the logistic model and the probit model, are considered with a more general mixture prior for genetic effects. Posteriors of marker effects can be estimated through Markov chain Monte Carlo (MCMC) techniques, where dimension reduction is effectively performed. Further statistical inference, not restricted to point estimation, can then be done based on the posteriors.

2.2 Model and Methods

2.2.1 Generalized linear models

Generalized linear models (GLMs)[65] provide a unified and flexible statistical framework for modeling how a response variable depends functionally on a set of predictors. Among GLMs, the logistic and probit regression models are used when the response variables are binary, as in our case, where the disease status is scored as affected or unaffected.

Suppose that there is a sample of N unrelated individuals from a single population, and genotypes of M SNPs for each individual (missing data is easily accommodated in this model). Individual i's phenotype can be connected with risk factors, such as SNP genotypes, by the logistic model

$$\text{logit}(P(Y_i = 1)) = \mu + X_i\beta, \quad i = 1, \cdots, N \tag{2.1}$$

or the probit model

$$\Phi^{-1}(P(Y_i = 1)) = \mu + X_i\beta, \quad i = 1, \cdots, N \tag{2.2}$$

where Y_i is individual i's binary phenotype, e.g. disease status of being affected ($Y_i = 1$) or unaffected ($Y_i = 0$), the intercept μ is the overall mean of the SNP effects, $X_i = (X_{i1}, \cdots, X_{iM})$ contains the genotype information of M SNPs for individual i, and β_j in the vector of $\beta = (\beta_1, \cdots, \beta_M)^T$ corresponds to the effect of the jth SNP. $X_i\beta$ takes the matrix multiplication which is equivalent to $\sum_{j=1}^{M} X_{ij}\beta_j$. The link functions are the logit function, $\text{logit}(\cdot)$, for model (2.1), and the inverse cumulative distribution function of the standard normal distribution, $\Phi^{-1}(\cdot)$, for model (2.2). Note that for now this formulation ignores non-genetic effects, and the effects of gene by gene and gene by environment interactions. This is purely to simplify notations, and these terms can readily be incorporated into our models as additional terms.

2.2.2 Genotype encoding

The genotypic effects is encoded using standard quantitative genetics models, so that for individual i and SNP j (or a two-allele gene) with alleles denoted A_j

and a_j, the element of the "design" matrix X is

$$X_{ij} = \begin{cases} 2, & \text{for genotype } A_jA_j \\ 1, & \text{for genotype } A_ja_j \\ 0, & \text{for genotype } a_ja_j \end{cases} \tag{2.3}$$

where only the additive effect is modeled. Alternatively, one can encode the genotype so that

$$\left(X_{ij}^{(add)}, X_{ij}^{(dom)}\right) = \begin{cases} (1,-1), & \text{for genotype } A_jA_j \\ (0,1), & \text{for genotype } A_ja_j \\ (-1,-1), & \text{for genotype } a_ja_j \end{cases} \tag{2.4}$$

where SNP j has its effects in the form of $\beta_j = \left(\beta_j^{(add)}, \beta_j^{(dom)}\right)$, where $\beta_j^{(add)}$ and $\beta_j^{(dom)}$ correspond to the additive effect and the dominance effect, respectively. One can also choose other ways of encoding effects.

2.2.3 Bayesian framework using hierarchical mixture priors

A Bayesian framework (Figure 2.1) is considered for the generalized linear models. Following Zhang *et al*. [63], we wish to classify each SNP effect under investigation into one of three classes: the positive-effect class, the negative-effect class, and the negligible-effect class. The number of SNP effects in each of the positive and the negative classes is likely to be small compared to the total number of SNPs, since it is expected that only a small number of SNPs are associated with the disease status. (Note that "positive" and "negative" refer to "liability" and "protective" allelic effects so that directionality is expected with regard to a reference allele at each SNP. We advocate consideration of directionality with regard to the "ancestral" vs. "derived" allele so that an implicit and consistent

evolutionary interpretation is carried through the analysis. However, this is not necessary, and the effect classes can be defined at the discretion of the investigator.)

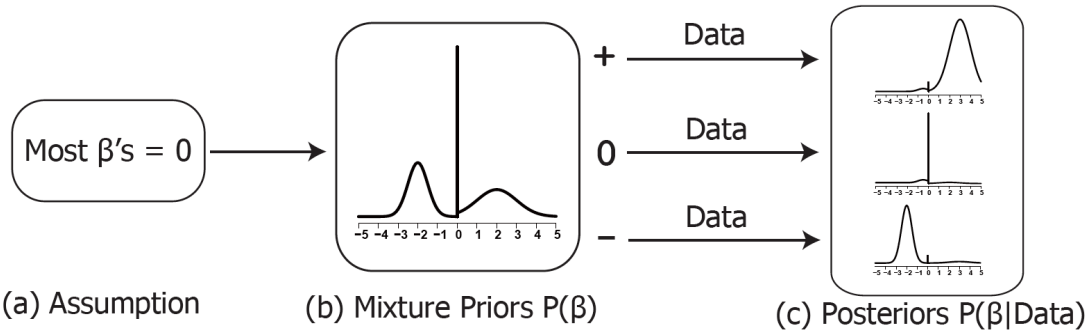

Figure 2.1: The Bayesian framework of mixture models. The Bayesian framework considers mixture priors for the SNP effects, which are derived from the assumption that most SNP effects are negligible. With the observation of data, the posterior distributions of the effects provide information of classifying the effects into the classes of positive, negative, or zero.

To develop such a model, I extend the idea of Zhang *et al.* [63] by using a 3-component mixture prior distribution for each effect β_j, irrespective of whether a logistic or probit model is considered:

$$\beta_j \overset{\text{i.i.d.}}{\sim} p_+ N_+(\mu_+, \sigma_+^2) + p_- N_-(\mu_-, \sigma_-^2) + p_0 I_{\{\beta_j = 0\}} \tag{2.5}$$

where p_+, p_- and $p_0 = 1 - p_+ - p_-$ are the probabilities for the effect to be classified into the positive-effect class, the negative-effect class, and the negligible class, respectively, $N_+(\mu_+, \sigma_+^2)$ and $N_-(\mu_-, \sigma_-^2)$ are normal distributions truncated at 0 with parameters $\mu_+, \sigma_+^2, \mu_-, \sigma_-^2$. The probability density functions of the two truncated normal distributions are the same as those in Zhang *et al.* [63]. The prior (2.5) is assumed to be independent among different β_j. In the case where $\beta_j = \left(\beta_j^{(add)}, \beta_j^{(dom)}\right)$ is employed, it is likely for additive and dominance effects to follow different distributions, so we can assume that there are different sets of

parameters $p_+, p_-, \mu_+, \sigma_+^2, \mu_-, \sigma_-^2$ for additive ($\beta_j^{(add)}$) and dominance ($\beta_j^{(dom)}$) effects, respectively.

The *a priori* information about the probability of a SNP with a positive, negative or negligible effect on the disease can be incorporated into the following Dirichlet prior distribution [63] for p_+ and p_-,

$$(p_+, p_-, 1 - p_+ - p_-) \sim \text{Dirichlet}\left(m_+^{(0)}, m_-^{(0)}, m_0^{(0)}\right) \tag{2.6}$$

Zhang *et al.* [63] set hyper-parameters $\left(m_+^{(0)}, m_-^{(0)}, m_0^{(0)}\right)$ to $(1, 1, 1)$, and assume p_+ and p_- are both restricted on the interval $[0, \min(m_0/M, 1)]$, where m_0 is a predefined parameter of order $\sqrt{N}$ (sample size) [64]. One can choose to follow this and assume the aforementioned truncated Dirichlet distribution for (2.6), or alternatively, one can use an unrestricted Dirichlet distribution where $m_+^{(0)}$ and $m_-^{(0)}$ are users' guess on the numbers of positive and negative effects, respectively, based on prior knowledge, and $m_0^{(0)} \triangleq M - m_+^{(0)} - m_-^{(0)}$.

Zhang *et al.* [63] assume $\mu_+ = 0$ and $\mu_- = 0$, and σ_+^2, σ_-^2 follow the prior distributions of

$$\sigma_+^2 \sim \text{Inv-}\chi^2(\nu_{+0}, \sigma_{+0}^2), \tag{2.7}$$

$$\sigma_-^2 \sim \text{Inv-}\chi^2(\nu_{-0}, \sigma_{-0}^2), \tag{2.8}$$

where Inv-$\chi^2(\cdot, \cdot)$ is the inverse χ^2 distribution [66]. To be more general, one can also impose hierarchical structures on the prior distributions of β and assume $\mu_+, \sigma_+^2, \mu_-, \sigma_-^2$ follow the prior distributions

$$\sigma_+^2 \sim \text{Inv-}\chi^2(\nu_{+0}, \sigma_{+0}^2), \quad \mu_+|\sigma_+^2 \sim N(\mu_{+0}, \sigma_+^2/\kappa_{+0}) \tag{2.9}$$

$$\sigma_-^2 \sim \text{Inv-}\chi^2(\nu_{-0}, \sigma_{-0}^2), \quad \mu_-|\sigma_-^2 \sim N(\mu_{-0}, \sigma_-^2/\kappa_{-0}) \tag{2.10}$$

where $\mu_{+0}, \kappa_{+0}, \nu_{+0}, \sigma_{+0}, \mu_{-0}, \kappa_{-0}, \nu_{-0}, \sigma_{-0}$ are the hyperparameters.

By using the hierarchical mixture prior (2.5) for each effect β_j in a Bayesian framework, those negligible effects tend to have a very high posterior probability at zero, and those effects that are significantly nonzero tend to have a very high posterior probability for being positive or negative (Figure 2.1). Therefore, the posterior distribution for each β can be utilized to classify the β's into the three classes.

2.2.4 Markov chain Monte Carlo for posterior simulation

Since both the logistic model (2.1) and the probit model (2.2) are popular and each has certain advantages over the other, the hierarchical mixture prior (2.5) can be applied to both models (2.1) and (2.2), which are henceforth called the logistic mixture model and the probit mixture model for simplicity. The goal is to obtain the posterior distribution for each effect β_j $(j = 1, \cdots, M)$. However, for both models, the posterior distribution of β_j can not be directly sampled from, necessitating the use of MCMC techniques. More specifically, I propose a Metropolis-Hastings algorithm as well as two Gibbs samplers [67] for the models. A Gibbs sampler recursively draws a sample for each parameter from its full conditional distribution, i.e. a distribution conditional on all other parameters. For these MCMC techniques, samples are retained until certain convergence criteria are satisfied.

When an MCMC algorithm is under construction, a parameter's full conditional distribution can be decided by its Markov blanket, which is defined as the set of variables composed of its parents, its offspring, and other parents of its offspring in a Bayesian network [68]. Conditional on its Markov blanket,

the parameter will be independent of all other parameters in the network, so its full conditional distribution only depends on all the variables in the Markov blanket. Figure 2.2 shows the Bayesian network representation of our models. An MCMC algorithm needs to draw a sample of the vector $\beta = (\beta_1, \beta_2, \cdots, \beta_M)^T$ from its full conditional distribution. In order to achieve that, write β_{-j} to denote the vector of β with β_j excluded, and an MCMC algorithm can draw a sample of each β_j, $(j = 1, 2, \cdots, M)$ from its full conditional distribution. Following the definition, each β_j's Markov blanket is actually the entire Bayesian network, which means the full conditional distribution depends on all other variables. After the distribution for each β_j is decided, a system scan or a random scan can be used to go over all the elements of β. In our implementation, the system scan is used for drawing β from its full conditional distribution.

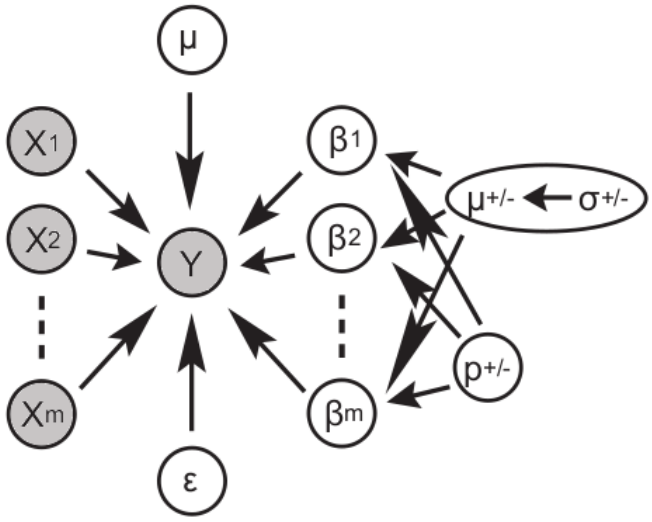

Figure 2.2: Graphical model and Markov blanket. The Markov blanket of each β_j includes β_j and all its parents, offspring, and other parents of its offspring. The observed variables are marked with filled nodes. This figure illustrates the Markov blanket for both the logistic and the probit models (some minor changes may be needed). MCMC algorithms can be constructed based on this figure.

At each iteration of the MCMC algorithm, each effect can be thought of as jumping stochastically among the three states of positive "+", negative "−", and negligible "0" effects (Figure 2.3). Conditional on the state, the value of the effect will follow a certain distribution. After a large number of iterations, the distribution of the effect will approach an equilibrium distribution, given by the posterior distribution of effect sizes conditional on the data.

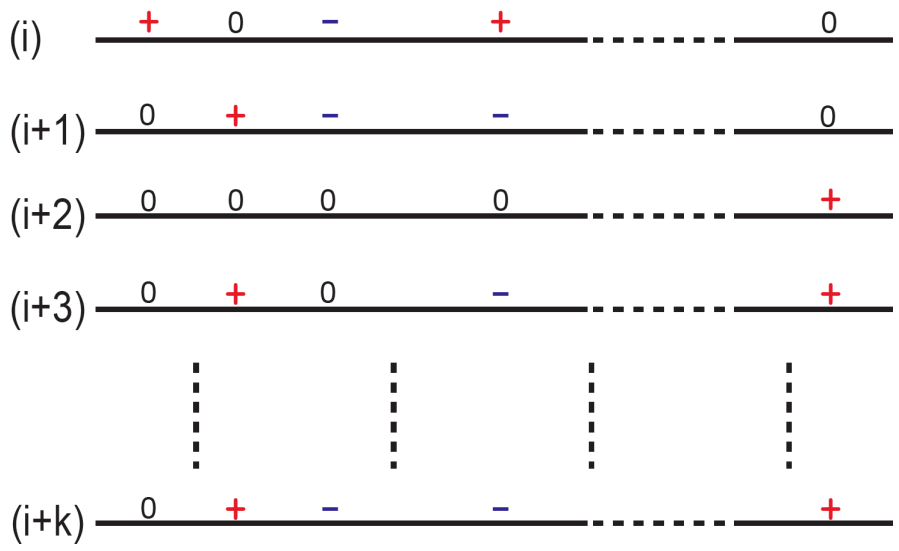

Figure 2.3: Illustration of steps for an MCMC algorithm. What is shown here is the iterations from step i to step $i+k$ for an MCMC algorithm. Each SNP effect stochastically jumps among the states of positive, negative and zero during the iterations.

Unlike classical hierarchical linear models where the mixture prior is conjugate [63], both the logistic and the probit models have a complicated full conditional distribution of β_j, which is difficult to sample from. As a common solution, some researchers have suggested using normal distributions to approximate the distribution by matching the mean and variance [66]. However, the approximation may be very rough and lead to non-negligible errors. To overcome the difficulty, I developed special techniques for both models in the following sections to draw β_j from its full conditional distribution. More specifically, I

have developed MCMC for the logistic mixture model using an adaptive independence sampler as well as Gibbs samplers using data augmentation methods (or the liability-threshold model) for both the logistic mixture model and the probit mixture model. The former method tries to solve the problem directly by using a Metropolis-Hastings algorithm to sample β_j, while the latter methods introduce an underlying liability (using a good approximation in the case of logistic) to transform the problem to one in the linear model.

Adaptive independence sampler for the logistic mixture model

Under the logistic mixture model, the full conditional distribution for β_j has a density proportional to

$$\frac{e^{-\beta_j \sum_i x_{ij}(1-y_i)}}{\prod_i (1+e^{-\mu-\sum_l \beta_l x_{il}})} \cdot \left\{ p_+ N_+(\beta_j|\mu_+, \sigma_+^2) + p_- N_-(\beta_j|\mu_-, \sigma_-^2) + p_0 I_{\{\beta_j=0\}} \right\} \tag{2.11}$$

where the notations are the same as in (2.5).

I propose using a special form of the Metropolis-Hastings algorithm, termed an adaptive independence sampler (AIS) [69], to sample β_j in the logistic mixture model. The proposal distribution is also a mixture normal distribution with the following form

$$q(\beta_j \to \beta_j^*) = q(\mathrm{sgn}(\beta_j), +)N_+(\beta_j^*|\mu_+, \sigma_+^2) + q(\mathrm{sgn}(\beta_j), -)N_-(\beta_j^*|\mu_-, \sigma_-^2) + q(\mathrm{sgn}(\beta_j), 0)I_{\{\beta_j^*=0\}} \tag{2.12}$$

where β_j^* is the proposed value and β_j is the current value. The function $q(s,t), \quad (s,t \in \{+,-,0\})$ is the pre-specified transition probability from state s to state t, and only depends on the signs of β_j and β_j^*, so the proposal distribution depends only on β_j's sign and is the so-called "independence sampler". The parameters μ_+, μ_-, σ_+^2 and σ_-^2, just as in the prior distribution of β_j (2.5), are

drawn from their full conditional distributions. The "prior" distribution hence changes at each iteration of the independence sampler, and the proposal distribution "adapts" to this change.

One example of the values for transition probability $q(s,t)$ is given in Figure 2.4. From the figure, it can observed that the proposal distribution will not draw zero for β_j if the current value is already zero, since otherwise β_j remains zero and there is no attempt to update at all. On the other hand, the proposal distribution will possibly draw a positive value, a negative value, or zero for β_j if the current value is nonzero, which guarantees the capacity of β_j to jump among the three possible classes. After obtaining a new value from the proposal distribution, one can follow the routine of the Metropolis-Hastings algorithm to compute the acceptance probability and decide whether to accept the proposed value or not.

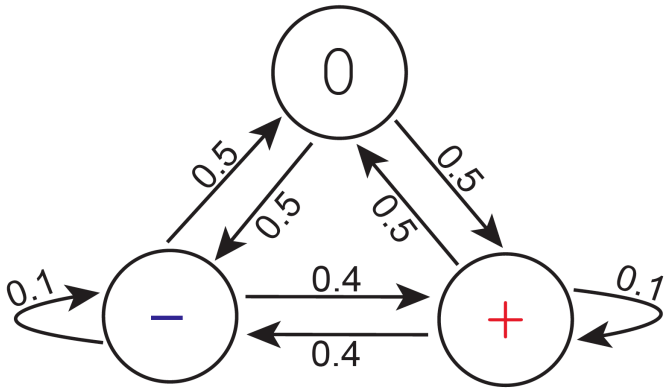

Figure 2.4: One choice of values for transition probabilities. Shown is one choice of the values for the transition probability $q(s,t)$ between any two states of positive, negative, and zero. The probability is zero for an effect to stay in the state of zero.

The advantage of using AIS is that it takes into account the changes in the prior distribution at different iterations: the proposal distribution (2.12) uses the same parameters μ_+, μ_-, σ_+^2, σ_-^2 as in the "prior" distribution (2.5). It makes

the proposal distribution a natural approximation for the posterior and helps obtain a satisfying acceptance rate of the Metropolis-Hastings algorithm. In our simulations, the transition probabilities $q(s,t)$ are kept the same for all j's and constant across all the iterations. As a possible improvement for future work, $q(s,t)$ can be extended to vary with j and the iterations of the MCMC to optimize the acceptance rate. For example, if some certain β_j is likely to be positive, the transition probability to propose a positive value can be larger, which can increase the acceptance rate.

The full conditional distribution for β_j (2.11) depends on all other parameters including β_{-j}. In other words, the MCMC regresses on the residuals of the model, which includes multiple effects at each iteration. It is different in principle from single marker analysis where only one effect is included in the model. By regressing on the residuals, the MCMC is expected to gain more power than single marker analysis.

Data augmentation for the probit mixture model

An alternative way is to use data augmentation techniques, which is mathematically equivalent to the liability-threshold model in genetics [70]. Since the link function of the model (2.2) is the inverse cumulative distribution function, no additional approximation is necessary. This method assumes that the binary variable Y_i is controlled by a latent continuous liability variable Z_i for individual i independently. Suppose $Y_i = I_{\{Z_i>0\}}$, then

$$\Phi^{-1}\left(P\left(Y_i=1\right)\right)=\Phi^{-1}\left(P\left(Z_i-\mu-X_i\beta>-\mu-X_i\beta\right)\right)=\Phi^{-1}\left(\Phi\left(\mu+X_i\beta\right)\right)=\mu+X_i\beta$$

holds if $Z_i \sim N(\mu + X_i\beta, 1)$. Equivalently,

$$Z_i = \mu + X_i\beta + \epsilon_i, \quad i = 1, \cdots, N$$

where $\epsilon_i \sim N(0, 1)$. The conditional distribution of Z_i given Y_i and other parameters is

$$Z_i|Y_i, \mu, X_i, \beta \sim N_+(\mu + X_i\beta, 1) I_{\{Y_i=1\}} + N_-(\mu + X_i\beta, 1) I_{\{Y_i=0\}} \tag{2.13}$$

The full conditional distribution of β_j $(j = 1, \cdots, M)$ can also be easily derived. With only moderate modification of the Gibbs sampler in Zhang *et al.* [63], I constructed a Gibbs sampler for the probit mixture model.

Data augmentation for the logistic mixture model

Similarly, the data augmentation technique can also be applied to the logistic mixture model. In a relevant study, Kinney and Dunson [71] use SSVS for logistic regression to select fixed effects as well as random effects. As for our purpose, since only fixed effects are considered, I use the similar techniques in their paper for our problem. What is distinctive here is that I use a three-component mixture prior as opposed to SSVS, and that I consider a much higher dimension than what they considered in their paper.

Independently for individual i, the binary variable Y_i is assumed to be controlled by a latent continuous liability variable Z_i. Suppose $Y_i = I_{\{Z_i>0\}}$, then $\text{logit}(P(Y_i = 1)) = \text{logit}(P(Z_i > 0)) = \mu + X_i\beta$ holds if Z_i follows a logistic distribution with the location parameter of $\mu + X_i\beta$, i.e.

$$P(Z_i \leq 0) = \frac{1}{1 + e^{-(z_i-\mu-X_i\beta)}} \tag{2.14}$$

Equivalently,

$$Z_i = \mu + X_i\beta + \xi_i, \quad i = 1, \cdots, N$$

where ξ_i follows a standard logistic distribution (with the location parameter of 0).

It is known that the standard logistic distribution can be well approximated by a t-distribution $t(\mu_t, \sigma_t, \nu)$ with the location $\mu_t = 0$, the degrees of freedom $\nu = 7.3$, and scaled by $\sigma_t = \sqrt{\frac{\pi^2}{3}\frac{\nu-2}{\nu}}$ [71]. The density function of the t-distribution $t(\mu_t, \sigma_t, \nu)$ is [72]

$$f(t;\ \mu_t, \sigma_t, \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu\sigma_t}}\left(1 + \frac{(t-\mu_t)^2}{\nu\sigma_t^2}\right)^{-\frac{\nu+1}{2}} \tag{2.15}$$

Figure 2.5 shows the curves of the standard logistic density as well as the density of $t(\mu_t, \sigma_t, \nu)$, suggesting that the two curves match each other very well.

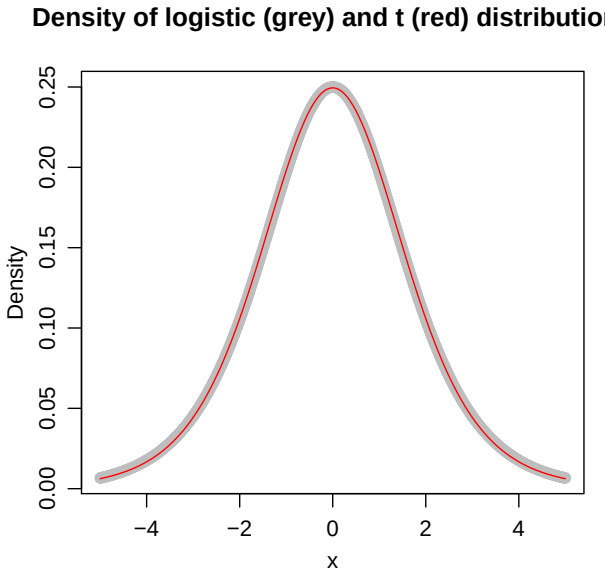

Figure 2.5: The density functions of two distributions. The grey curve is the standard logistic distribution, and the red one is the t-distribution $t(\mu_t, \sigma_t, \nu)$.

Moreover, marginally $\epsilon_i \sim t(\mu_t, \sigma_t, \nu)$ if

$$\epsilon_i|\phi \sim N\left(\mu_t, \phi^{-1}\right) \quad \text{and} \quad \phi \sim \Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\sigma_t^2\right)$$

where the density function of $\Gamma(a,b)$ is $f(x;\ a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} I_{\{x>0\}}$ [72]. Therefore, Z_i can be approximated by a linear combination of predictors and a normally distributed error term, i.e.

$$Z_i \approx \mu + X_i\beta + \epsilon_i, \quad i = 1, \cdots, N \tag{2.16}$$

where ϵ_i is as described previously. Upon using this approximation, we now have a linear model with a normally distributed liability as the response. The conditional distribution of Z_i given Y_i and other parameters is

$$Z_i|Y_i, \mu, X_i, \beta, \phi \sim N_+\left(\mu + X_i\beta, \phi^{-1}\right) I_{\{Y_i=1\}} + N_-\left(\mu + X_i\beta, \phi^{-1}\right) I_{\{Y_i=0\}} \tag{2.17}$$

and the full conditional distribution of β_j $(j = 1, \cdots, M)$ can now be easily derived. With only moderate modification of the Gibbs sampler in Zhang *et al.* [63], a Gibbs sampler was constructed for the logistic mixture model.

2.2.5 Heuristic methods for screening SNPs within the MCMC

MCMCs need to run many iterations before reaching equilibrium in polynomial computational time with respect to the number of SNPs. Current GWA studies often investigate a large number of SNPs ($> 1,000,000$), so that running MCMCs on these studies may be computationally costly. Under the assumption that most of the effects are expected to be insignificant and negligible, I propose that it is likely not necessary to update all the variables in the MCMCs. Here I develop a heuristic method, termed the score averaging method, to rule out a large number of effects using short runs of the MCMC. I apply the score averaging method within certain iterations during the "burn-in" period, and call it a "hierarchical-burn-in" period. After a certain number of effects have been

ruled out, the MCMC enters the "burn-in" period, and samples can be retained for estimating the posterior distribution after that.

The proposed score averaging is based on the score test statistic. Suppose $l(\beta_j; k)$ is the log-likelihood function given the values of all other parameters at the end of the iteration step k, and is treated as a function of β_j. Define the score function S_j and the information function I_j as

$$S_j \triangleq \frac{\partial l(\beta_j; k)}{\partial \beta_j}|_{\beta_j=0}, \quad I_j \triangleq -\frac{\partial^2 l(\beta_j; k)}{\partial \beta_j^2}|_{\beta_j=0}$$

Let C_j be an indicator variable as to whether β_j is to be updated. For each j still under investigation ($C_j = 1$), the method computes the running averages $\bar{S}_j^{(k)}$ of S_j^2/I_j over all the available k steps of iterations. At the end of the "hierarchical-burn-in" period, a user-defined number of effects with smallest running averages are excluded from the investigation in the successive iterations. The details of the procedure in the "hierarchical-burn-in" period are described below:

1. $k = 0$, $\bar{S}_j^{(k)} = 0$, $C_j = 1$, $(j = 1, 2, \cdots, M)$
2. Update $k = k + 1$. At the end of iteration k, compute S_j and I_j, and

$$\bar{S}_j^{(k)} = \frac{k-1}{k}\bar{S}_j^{(k)} + \frac{1}{k}S_j^2/I_j$$

   for each $j \in \{j : C_j = 1\}$
3. If $k < L$, then go to (2); otherwise, sort $\bar{S}_j^{(k)}$ for all $j \in \{j : C_j = 1\}$ and set $C_j = 0$ for those j's corresponding to the first M_{drop} smallest $\bar{S}_j^{(k)}$.

where L is the number of iterations for computing the running averages, and M_{drop} is the number of SNPs to be excluded at the end of L iterations. One may also want to employ this method several times to gradually reduce the number of SNPs to be considered in the model.

The goal of the score averaging is to significantly save future computation time on those SNPs with negligible effects. Although there is no theoretical support currently for the length of the "hierarchical-burn-in" period and the number of effects to be removed, I suggest applying this heuristic method to multiple chains for comparisons and using as many iterations as possible.

The rationale for the method arises from the score test for the null hypothesis of $\beta_j = 0$. (In principle one can also consider the likelihood ratio test or the Wald test). The advantage of the score test is that it uses only the slope and curvature information of $l(\beta)$ at $\beta = 0$, while the other two tests require the maximum likelihood estimate $\hat{\beta}$, which demands many more computations.

2.2.6 Missing genotypes

An issue often encountered in practice is missing information in genotype data. By treating genotypes as random, the MCMC algorithm provides a convenient way for imputation, which is to sample the missing genotypes from their full conditional distributions: for X_{ij} that is missing,

$$X_{ij}|\text{Data} \sim \text{Likelihood} \times P(X_{ij}|X_{i1}, \cdots, X_{i,j-1}, X_{i,j+1}, \cdots, X_{im})$$

The conditional distribution $P(X_{ij}|X_{i1}, \cdots, X_{i,j-1}, X_{i,j+1}, \cdots, X_{im})$ can be approximated by the dstribution $P(X_{ij}|Y_i)$, or by estimates of the genotype frequencies from some phasing software such as fastPHASE [73]. These approximations would be preferable, especially when the sample size or the number of markers is large, or the missing rate of genotypes is relatively high.

2.3 Simulation Results

For evaluation, the performance of our proposed methods is compared to that of single marker analysis on simulated data sets. Since the data augmentation for the logistic mixture model was developed more recently, a different set of simulations is considered for it from those for the other two methods. All methods are expected to have similar performance (as shown in the results), and different scenarios are used to evaluate the performance from different aspects.

2.3.1 Simulations for the logistic model with AIS and the probit model

Instead of using samples generated by a coalescent-based approach [74], I use the samples from the GSK-POPRES project [75] which consists of Affymetrix 500K SNP chip data on thousands of individuals with various ethnic background. Three kinds of methods have been applied to each data set.

1. Single marker analysis: Each SNP was tested independently by a genotype test with 2 degrees of freedom.
2. Mixture models with score averaging: Both the logistic and probit models were applied, with score averaging acceleration for the MCMCs. The score averaging method gradually reduced the numbers of additive effects and dominance effects to 50, respectively, in the "hierarchical burn-in" period of MCMCs.
3. Mixture models without score averaging: 100 SNPs with the smallest p-

values from genotype tests were chosen, and the logistic and probit models were applied to these SNPs directly. No score averaging acceleration was used.

GSK-POPRES Affymetrix SNP Data

We considered 1,115 Swiss individuals genotyped on the Affymetrix 500K SNP chip as part of the GSK-POPRES, and only the SNPs on the first chromosome are studied. SNPs are selected by including only high-quality genotypes regarding allele frequencies and missingness. Those SNPs on the 1st chromosome are filtered using the criterion that minor allele frequencies (MAF) above 1%, missing genotypes per individual less than 10%, missing genotypes per SNP less than 10%, and the p-values for the Hardy-Weinberg Equilibrium are at least 0.001. After the filtering, 28,880 SNPs satisfying the above criterion are considered for the simulations. The physical locations of the SNPs on the chromosome vary from 742,429 bp to 247,134,313 bp, i.e. in the range of approximately 247 Mb.

Data Simulation

In order to simulate disease susceptibility loci (DSLs), all the SNPs with completely observed genotypes are first extracted. There are 2,339 such SNPs with locations from 1,120,590 bp to 247,040,508 bp providing good coverage of all the SNPs on the first chromosome. Fifty groups of SNPs are randomly selected to be DSLs with each group including 10 DSLs. The 500 DSLs correspond to 441 distinct SNPs. In each of the 50 groups, the smallest distance of flanking DSLs is 6,826,456 bp, suggesting the DSLs are effectively unlinked.

The effect sizes for our simulations depend on the way genotypic effects are encoded. If the encoding in (2.4) is used to model both the additive effect and the dominance effect, the contribution of a SNP encoded by (X_1, X_2) can be formulated as $X_1\beta_1 + X_2\beta_2$, where β_1 and β_2 are the additive effect and the dominance effect, respectively. In the logistic model, e^{β_1} and e^{β_2} have the natural interpretations as heterozygote odds ratio and homozygote odds ratio.

The absolute values of the effects are sampled from the Gamma ($\Gamma(\cdot, \cdot)$) distributions: for each DSL, the additive effect satisfies $|\beta_1| \sim \Gamma(3, 1)$ truncated at [1,4], and the dominance effect satisfies $|\beta_2| \sim \Gamma(2, 1)$ truncated at [1,3]. The sign of each effect is chosen to be positive or negative with equal probabilities.

To simulate phenotypes, an error term ϵ_i is added to the linear combination $\sum_j X_{ij}\beta_j$ for individual i , where $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ independently for all i. The variance σ_ϵ^2 is chosen such that the variance of $\sum_j X_{ij}\beta_j$ is 100% or 90% of the total variance of $\sum_j X_{ij}\beta_j + \epsilon_i$, providing approximately a targeted narrow sense heritability of 100% or 90% . Then the error terms ϵ_i's are sampled from $N(0, \sigma_\epsilon^2)$. The value of μ is selected such that the linear combinations $\mu + \sum_j X_{ij}\beta_j + \epsilon_i$ have 0 as the upper 20% percentile. The phenotype Y_i can be sampled according to the probability $P(Y_i = 1)$, which can be computed via the logistic model or the probit model. The end result of this collection of parameters is to produce genotype and phenotype data equivalent to certain narrow sense heritability, in reasonable agreement with that of complex phenotypes of medical interest.

After the phenotypes are simulated, 200 cases and 200 controls are randomly sampled from the 1,115 individuals, and these 400 selected individuals are considered the samples of a case-control study. The prevalence of the disease is approximately 0.2.

For each combination of a mixture model (logistic or probit) and an error term (without or with), 20 case-control data sets are simulated using the above procedures, each data set is simulated from 10 DSLs and contains 200 cases and 200 controls.

Evaluation of Simulation Results

One realization of the results of the logistic mixture model is shown in Figure 2.6. The top panel gives the minus logarithm of p-values for single marker genotype tests, while the bottom panel shows the posterior probabilities of effects being nonzero given by the logistic mixture models. The figure shows that, at least for this specific case, single marker analysis will miss some DSLs, which the logistic mixture model can identify.

Precision-Recall (PR) curves are used here to compare the mixture model methods with single marker analysis. A Precision-Recall curve is said to dominate another PR curve if the former curve is closer to the upper-right corner of the figure. Similar to the Receiver Operator Characteristic (ROC) curves that are commonly used as a performance evaluation tool [63], Precision-Recall curves are also often used, especially when dealing with highly skewed data sets. A deep connection has been shown between the ROC space and the PR space that a curve dominates in the ROC space if and only if it dominates in the PR space [76]. Since the number of DSLs and their associated markers is much smaller than the number of markers not associated with any DSL, we use a PR curve instead of a ROC curve for this evaluation.

Since the goal of our analysis is to identify multiple DSLs, the number of

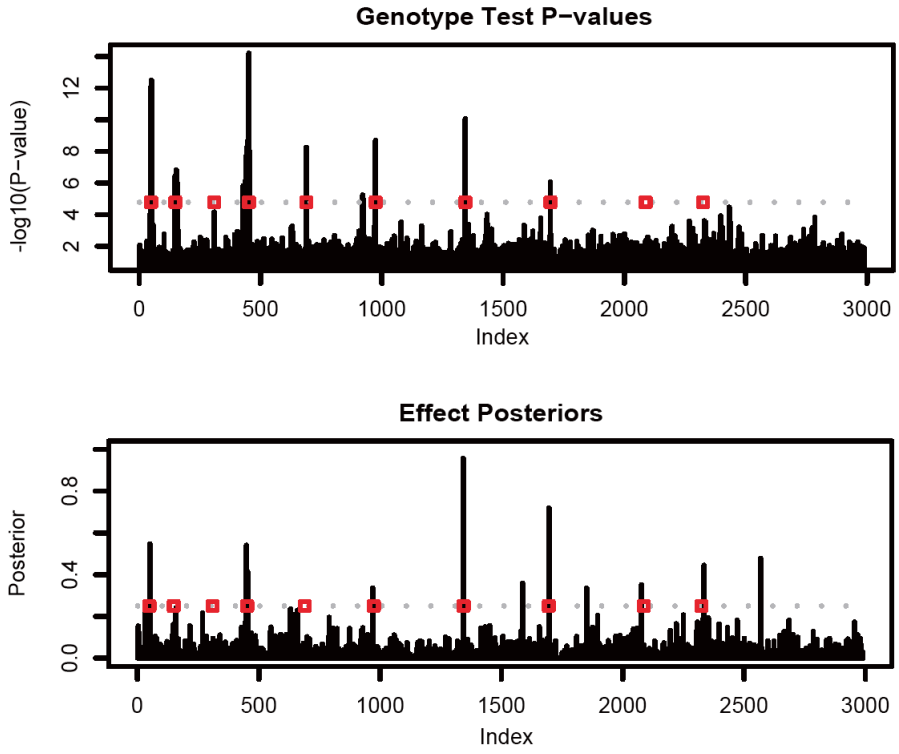

Figure 2.6: One instance of simulation results. The results are obtained by using single marker analysis and the logistic mixture model, with both methods applied to the same data set. The upper panel gives a Manhattan plot for the genotype tests, and the lower panel gives the posterior probability for each SNP to have a nonzero additive effect or a nonzero dominance effect.

DSLs that can be captured is more informative for evaluation than the number of claimed significant SNPs that are in linkage disequilibrium with DSLs. In this chapter, a DSL is identified, if at least one significant SNP is associated with the causal SNP (e.g. $r^2 > 0.5$). Based on this definition, we define the precision and the recall as follows:

$$\text{Precision} \triangleq \frac{\#(\text{identified DSLs})}{\#(\text{claimed significant SNPs})} \qquad \text{Recall} \triangleq \frac{\#(\text{identified DSLs})}{\#(\text{true DSLs})}$$

Now we can compare the performance of different methods: if the PR curve of a method dominates the one of another method, the former method outperforms the latter one. Similar to ROC curves where the area under curves can be

compared, the larger the area under the PR curve, the better performance the method has.

1) PR curves for data sets assuming no error terms

We first examined the performance between SMA and mixture models with score averaging. Figure 2.7 gives the Precision-Recall curves on case-control data sets without random errors. It shows that for both the logistic and probit models, most solid curves are closer to the upper-right corner, i.e. dominates the dashed curves from SMA of the same data set. The figure shows that the Bayesian mixture models nearly always outperform SMA at least for this specific simulation scenario.

We further studied the area under each of the precision-recall curves in Figure 2.8.

We can observe that the areas under the PR curves of the logistic mixture model are comparable with those of the probit mixture model, and both are the larger than those of SMA. When pairwise differences of the areas under the curves are considered, the differences between the logistic mixture model (with score averaging) and single marker analysis have a mean of 0.2369, a standard deviation of 0.0979, and the 25%, 50%, 75% quantiles are 0.1507, 0.2440, 0.2975, respectively; the differences between the probit mixture model (with score averaging) and single marker analysis have a mean of 0.2189, a standard deviation of 0.1096, and the 25%, 50%, 75% quantiles are 0.1106, 0.2076, 0.3134, respectively. These simulation results suggest that our proposed approach is nearly always more powerful than SMA.

Within the mixture models, the areas for models with score averaging are

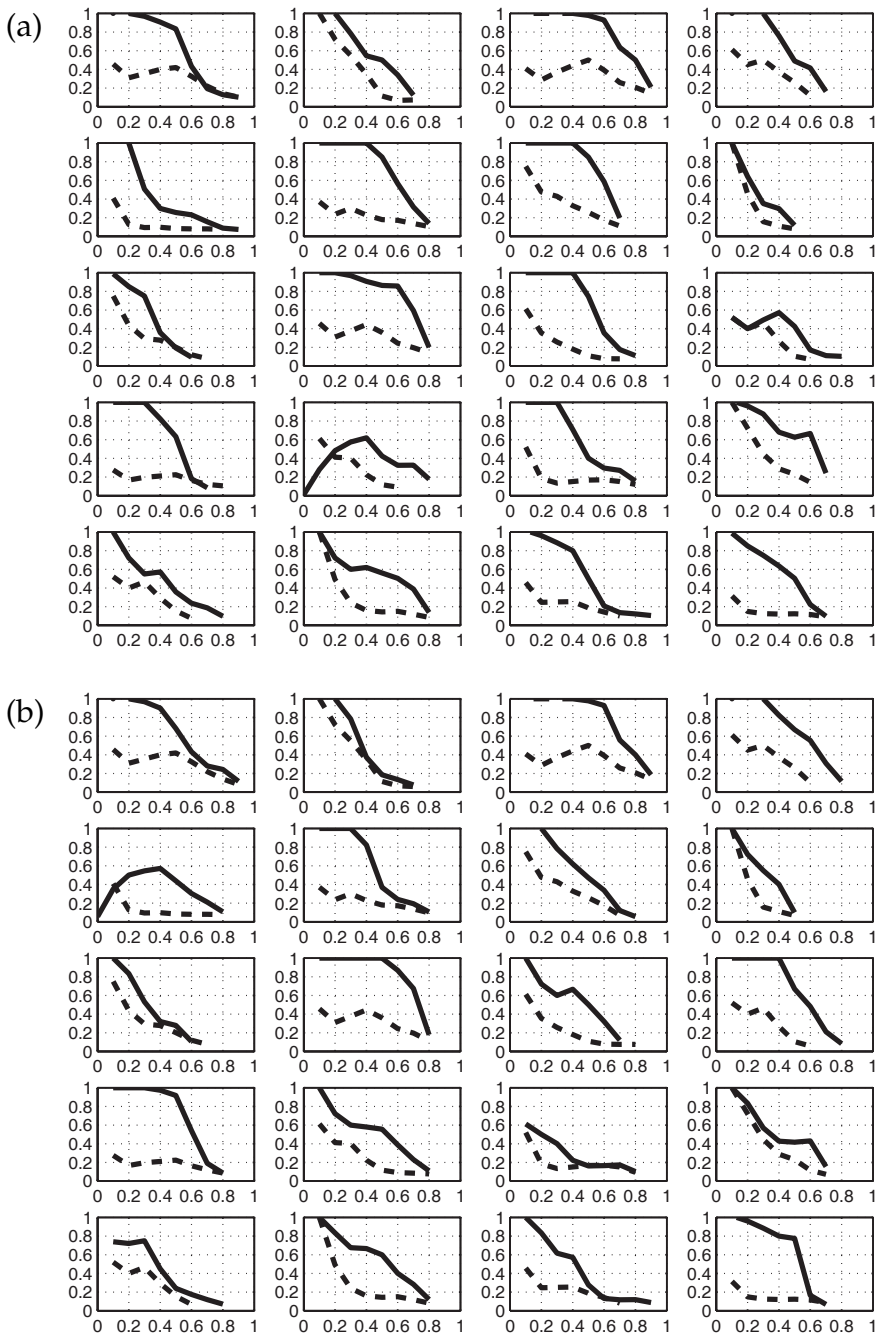

Figure 2.7: Precision-recall curves for data sets without random errors. The y-axes denote precision and the x-axes denote recall. Each curve, smoothed using lowess, is based on one of 20 case-control data sets. Panel (a) corresponds to the results of the logistic mixture model (solid lines) and single marker analysis (dashed lines); panel (b) shows the results of the probit mixture model (solid lines) and single marker analysis (dashed lines).

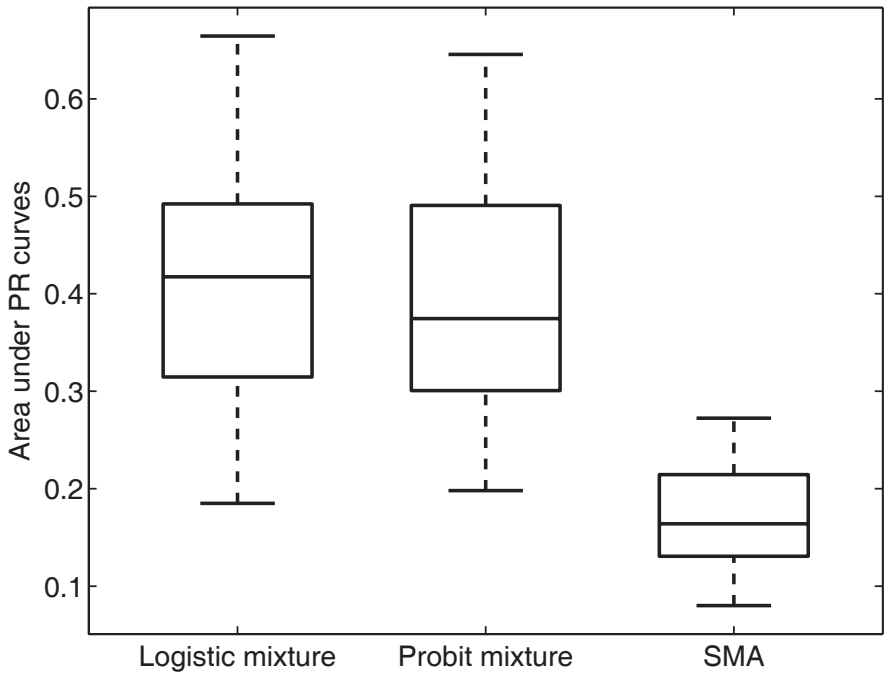

Figure 2.8: Box plots of areas under PR curves for data sets without random errors. Shown are the box plots of the area under the precision-recall curves for the logistic mixture model, the probit mixture model, and single marker analysis on the 20 case-control data sets without random errors.

larger than those without using score averaging (not shown here). It suggests that running mixture models with acceleration on all SNPs can be more powerful than running the models directly on a small set of SNPs, although the latter can save some time. The proposed score averaging method provides researchers with a choice as the trade-off between speed and power.

2) PR curves for data sets assuming error terms

For the data sets assuming error terms in the linear combination $\mu + \sum_j X_{ij}\beta_j + \epsilon_i$, PR curves were analyzed in a similar way. Figures 2.9 and 2.10 show the Precision-Recall curves on case-control data sets with random errors as well as the areas under the curves.

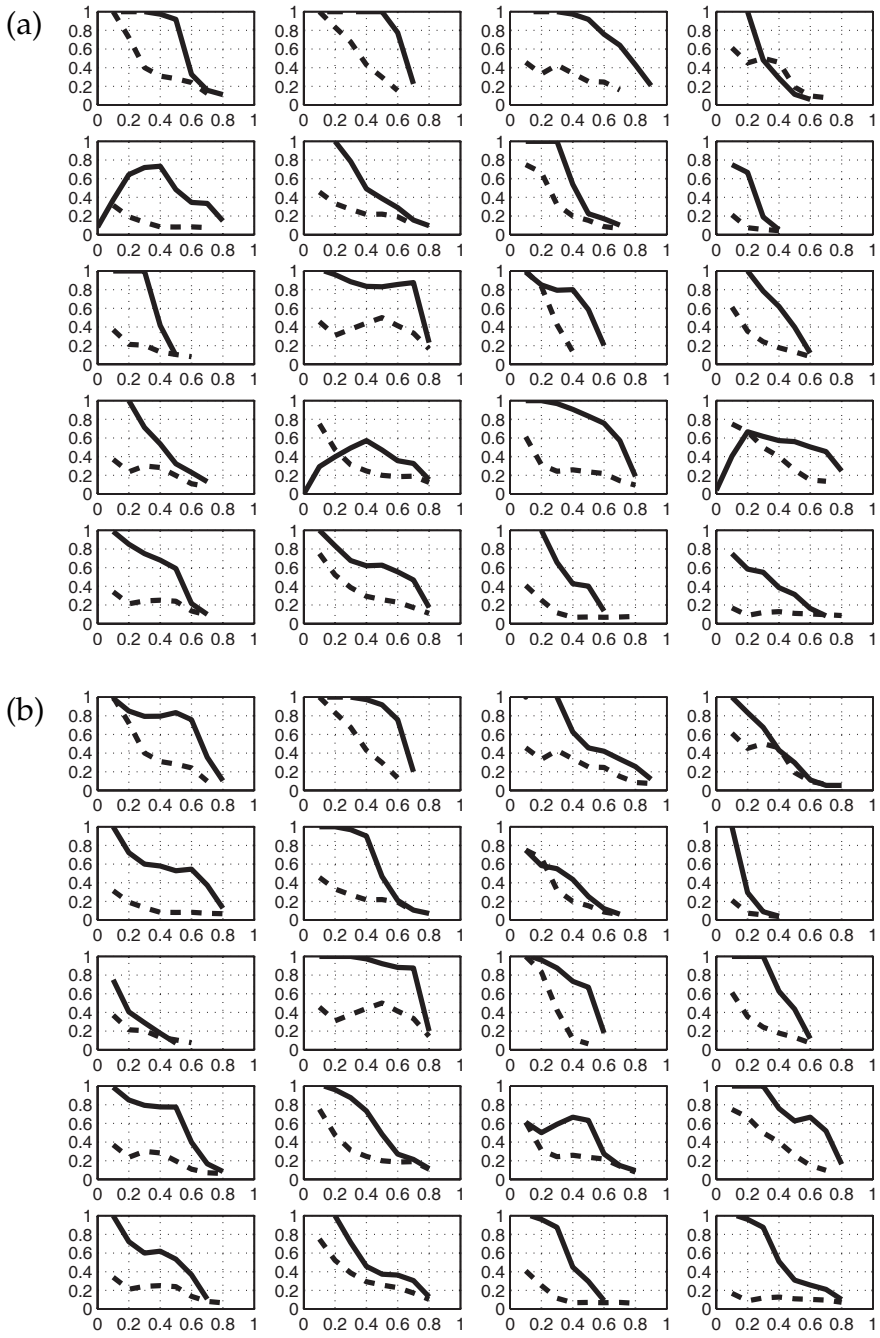

Figure 2.9: Precision-recall curves for data sets with random errors. The y-axes denote precision and the x-axes denote recall. Each curve, smoothed using lowess, is based on one of 20 case-control data sets with random errors. Panel (a) corresponds to the results of the logistic mixture model (solid lines) and single marker analysis (dashed lines); panel (b) shows the results of the probit mixture model (solid lines) and single marker analysis (dashed lines).

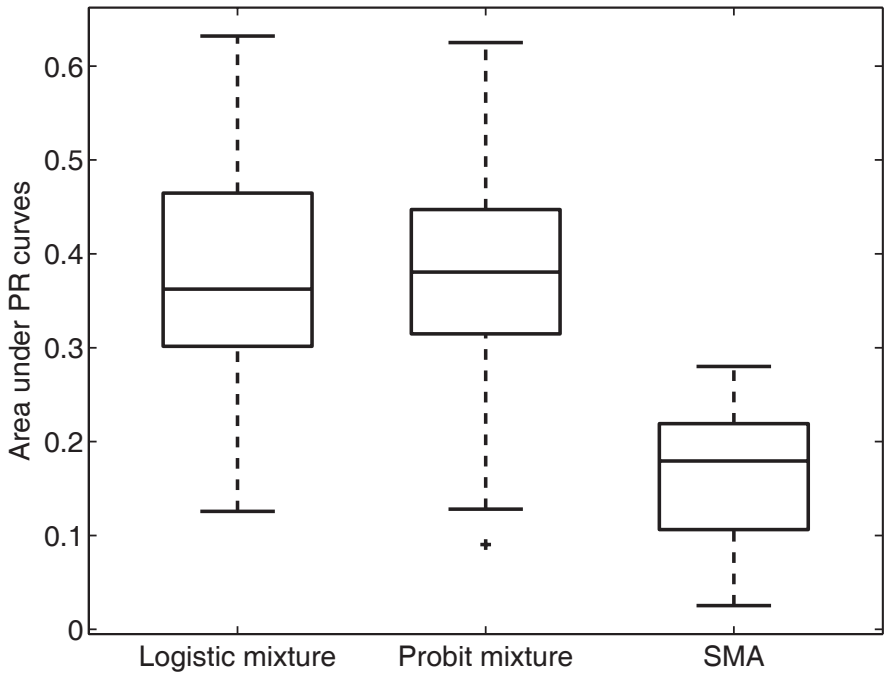

Figure 2.10: Box plots of areas under PR curves for data sets with random errors. Shown are the box plots of the area under the precision-recall curves for the logistic mixture model, the probit mixture model, and single marker analysis on the 20 case-control data sets with random errors.

The results are similar to those in Figures 2.7 and 2.8. When pairwise differences of the areas under the curves are considered, the differences between the logistic mixture model (with score averaging) and single marker analysis have a mean of 0.2145, a standard deviation of 0.0925, and the 25%, 50%, 75% quantiles are 0.1650, 0.2099, 0.2507, respectively; the differences between the probit mixture model (with score averaging) and single marker analysis have a mean of 0.2067, a standard deviation of 0.0904, and the 25%, 50%, 75% quantiles are 0.1447, 0.2314, 0.2718, respectively. The results suggest that our proposed approach is robustly more powerful than SMA even when a portion of the variation in disease is not attributable to genetic factors.

2.3.2 Simulations for the logistic model with data augmentation

In the simulations to evaluate the logistic mixture model with data augmentation, we compare its performance to that of SMA in a special case. Specifically, the SNPs are assumed to be independent of each other, and suppose from prior knowledge all causative alleles are known to be minor alleles. SMA can not take into account such prior information, but the logistic mixture model using the 3-component prior (2.5) is able to model this through small probabilities p_-. For convenience, we use GBLOG (Gibbs sampler for Bayesian LOGistic model) to denote the logistic model with data augmentation in this part. The two methods, GBLOG and SMA, are assessed by the ROC curves. A curve closer to the upper left corner indicates a better method. Alternatively, we can also look at the ranks assigned to the true causative SNPs. On the one hand, SNPs can be ranked by GBLOG in a descending order of $P\left(\beta_j > 0\right) + P\left(\beta_j < 0\right)$, which is the posterior probability that the SNP effect is nonzero; on the other hand, SNPs can be ranked by SMA in an ascending order of p-values. The method giving higher ranks for the true causative SNPs is preferable. As an example, Figure 2.11 shows a specific case that GBLOG outperforms SMA in the sense of giving higher ranks, i.e. $Rank_{SMA} - Rank_{GBLOG} > 0$, for causative SNPs, where $Rank_{SMA}$ and $Rank_{GBLOG}$ are the ranks given by the two methods for a causative SNP.

Five different scenarios A-E are considered with different odds ratios and allele frequencies of the causal SNPs (Table 2.1). Each of the scenarios considers 100 case-control studies with equal sample size of 1,000 (500 cases and 500 controls), and assumes 10 causal SNPs in each study. To simulate genotypes, scenarios A, B and C follow Zhang and Liu [77] by computing the causal allele

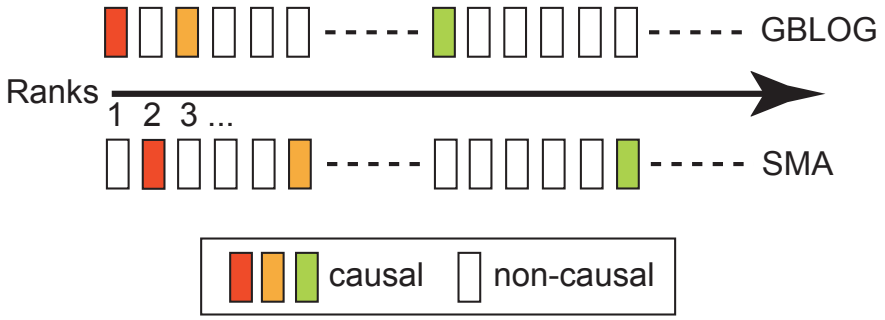

Figure 2.11: A special case that GBLOG outperforms SMA in the sense of giving higher ranks for causal SNPs.

frequencies in the cases using ORs in Table 2.1; while scenarios D and E use PLINK [78] to simulate the data sets, given the parameters in Table 2.1. After the data sets are simulated, both SMA and GBLOG are applied to the same data sets. For scenarios A and B ($M = 2,000$), GBLOG was run directly on all SNPs; for scenarios C ($M = 10,000$), D ($M = 10,010$) and E($M = 100,010$); for scenarios C, D and E, prior to using GBLOG, screening was first applied either to pick 2,000 SNPs with the smallest SMA p-values (for C), or to pick SNPs with p-values$\leq$0.05 leaving $\sim$ 500 for D, and $\sim$ 5000 for E.

ROC curves are plotted by coupling all the 100 studies in each scenario (Figures 2.12, 2.13 and 2.14): The p-values from SMA are put together and sorted in an ascending order, and the posteriors $P\left(\beta_j \neq 0\right)$ are put together across the studies and sorted in a descending order. We can observe that, in terms of ROC, GBLOG outperforms SMA especially for higher true positive rates. With higher ORs of the causal SNPs, both methods tend to have higher power in identifying the signals (Figures 2.12 and 2.13).

Table 2.1: Setup of the simulation scenarios. In each of the five scenarios, 10 causal SNPs are simulated, based on the causal allele frequencies (CAF) in controls and the heterozygous odds ratios (OR) shown in column 3-12, and certain number of non-causal SNPs are also simulated. Prevalence is given in scenarios D and E which use PLINK to simulate the data set.

Scenario		Causal SNPs										#non-causal	Prevalence
A	*CAF*	0.5	0.5	0.3	0.3	0.3	0.5	0.5	0.3	0.3	0.3	1,990	–
	OR	1.3	1.3	1.3	1.3	1.3	1.2	1.2	1.2	1.2	1.2		
B	*CAF*	0.2	0.2	0.1	0.1	0.1	0.2	0.2	0.1	0.1	0.1	1,990	–
	OR	1.2	1.2	1.2	1.2	1.2	1.1	1.1	1.1	1.1	1.1		
C	*CAF*	0.3	0.3	0.2	0.2	0.2	0.3	0.3	0.2	0.2	0.2	9,990	–
	OR	1.3	1.3	1.3	1.3	1.3	1.2	1.2	1.2	1.2	1.2		
D	*CAF*	0.3	0.3	0.2	0.2	0.3	0.3	0.3	0.2	0.2	0.2	10,000	0.1
	OR	1.5	1.5	1.5	1.5	1.3	1.3	1.3	1.3	1.3	1.3		
E	*CAF*	0.3	0.3	0.2	0.2	0.3	0.3	0.3	0.2	0.2	0.2	100,000	0.1
	OR	1.5	1.5	1.5	1.5	1.3	1.3	1.3	1.3	1.3	1.3		

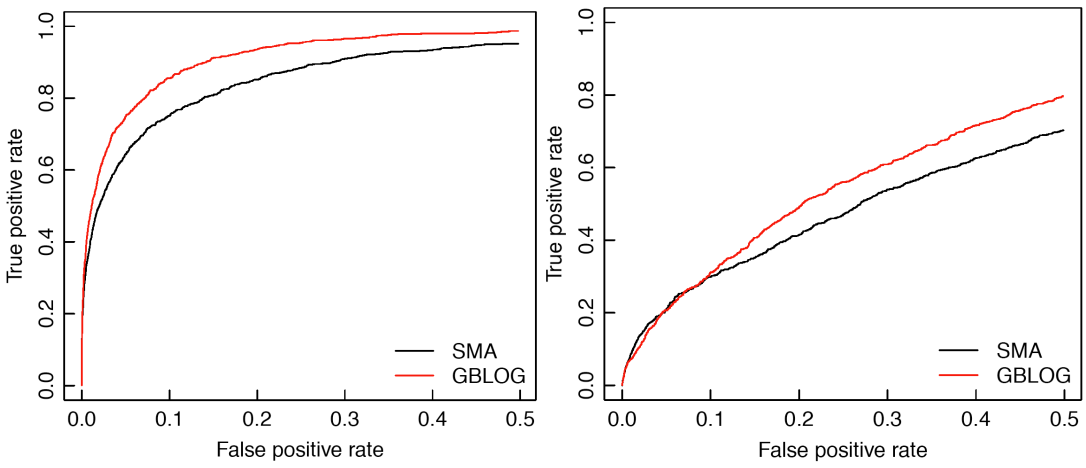

Figure 2.12: ROC curves of GBLOG and SMA in scenarios A (left panel) and B (right panel). Both scenarios simulate 2,000 SNPs including 10 causal ones.

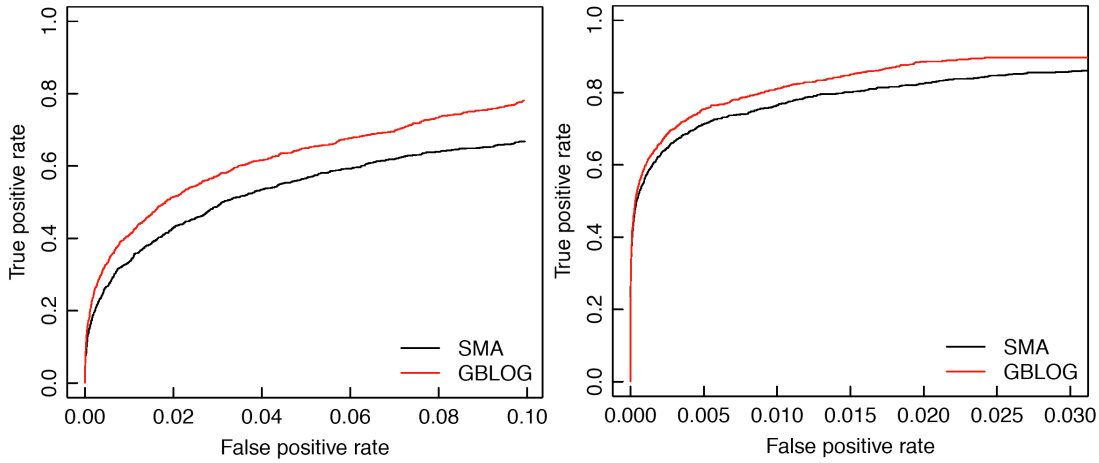

Figure 2.13: ROC curves of GBLOG and SMA in scenarios C (left panel) and D (right panel). Both scenarios simulate ~10,000 SNPs including 10 causal ones.

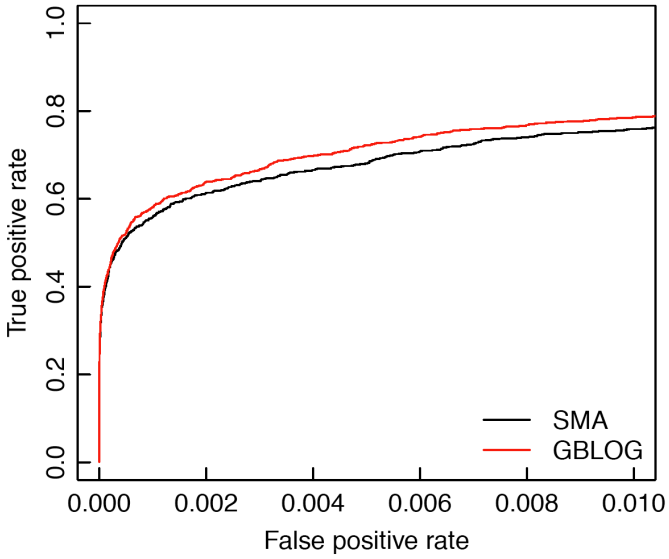

Figure 2.14: ROC curves of GBLOG and SMA in scenario E, which simulates 100,000 SNPs plus 10 causal ones in each study.

Ranks given by the two methods are also compared by looking at the difference $Rank_{SMA} - Rank_{GBLOG}$ for causal SNPs. As mentioned before, if a method gives higher ranks for most of the causal SNPs, then this method is thought to

have better performance. Figures 2.15 and 2.16 show the histograms and scatter plots of rank differences for scenario A and C, respectively. From the figures, we can observe that in most cases, GBLOG gives higher ranks to causal SNPs than SMA, and that the difference is large especially for the ones with smaller effects.

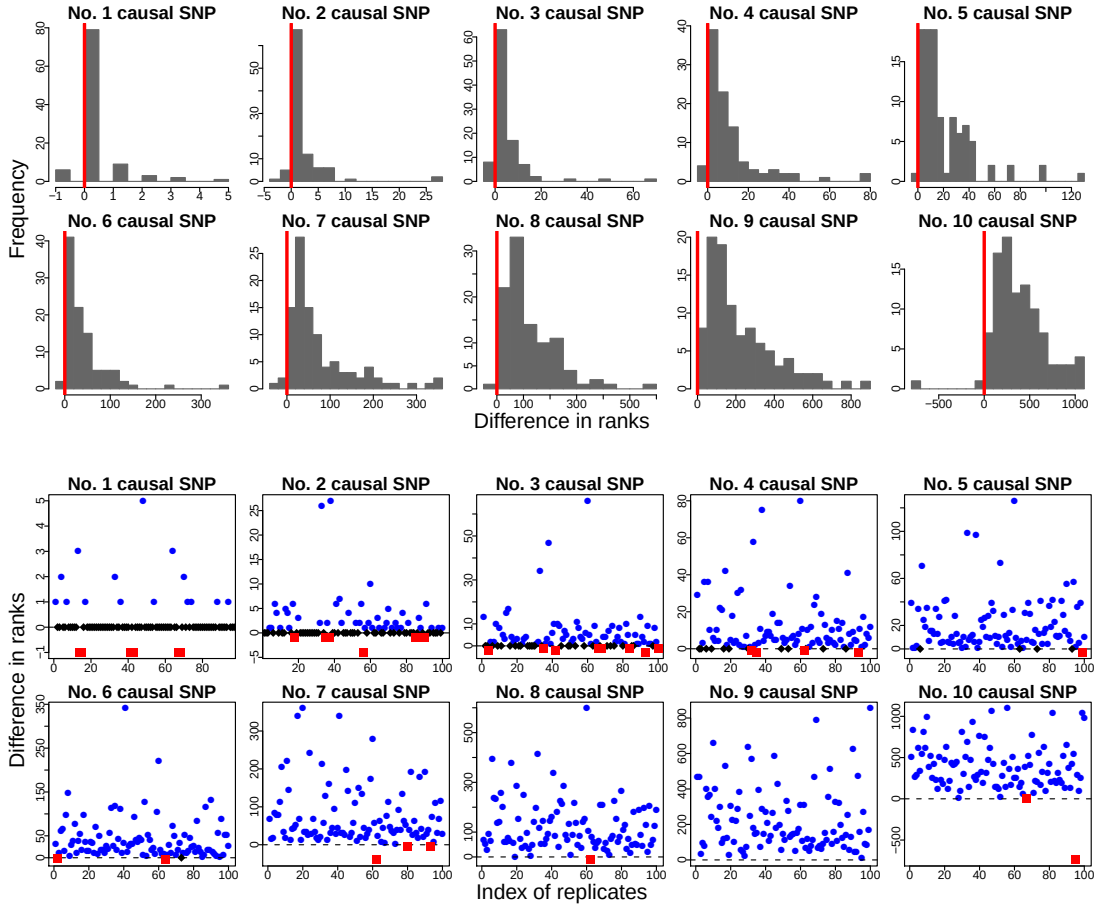

Figure 2.15: Rank comparisons of GBLOG and SMA in scenario A. The upper panel shows the histograms of rank differences for all 10 causal SNPs, and the lower panel shows the scatter plots of the differences. The dots in blue indicate positive differences ($Rank_{SMA} - Rank_{GBLOG} > 0$), and the squares in red indicate negative differences.

The three-component mixture prior (2.5) explicitly models the probabilities of effects being positive and negative in p_+ and p_-, makes it useful when prior

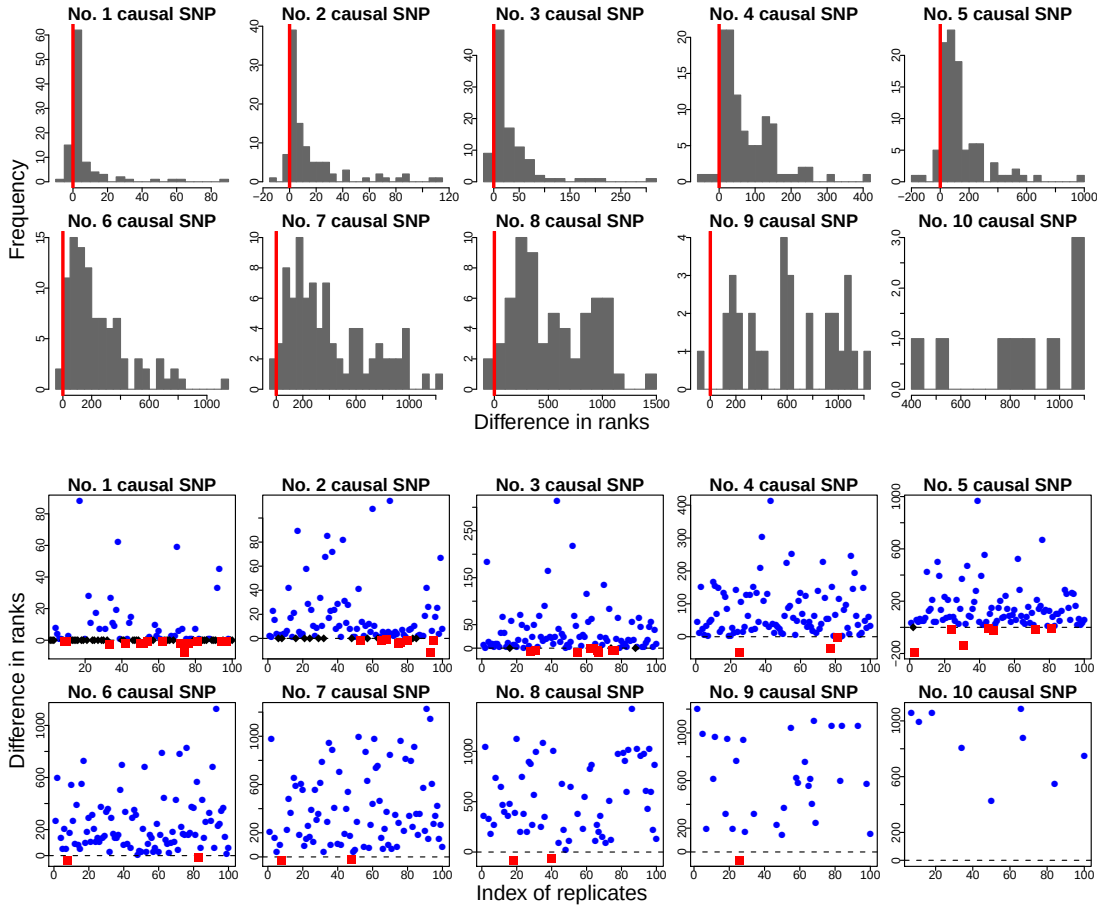

Figure 2.16: Rank comparisons of GBLOG and SMA in scenario C. The upper panel shows the histograms of rank differences for all 10 causal SNPs, and the lower panel shows the scatter plots of the differences. The dots in blue indicate positive differences ($Rank_{SMA} - Rank_{GBLOG} > 0$), and the squares in red indicate negative differences.

information is available. In the above simulations, all causal alleles are assumed to be minor ones. Even though all SNPs are simulated to be unlinked and no LD is assumed, GBLOG still outperforms SMA in the sense of elevating ranks for causal SNPs, and this gain is likely due to the fact that GBLOG is able to take into account the prior knowledge. So this set of simulations suggest that using the 3-component mixture prior can be advantageous in some cases.

2.4 Discussion

A fundamental challenge in genome-wide association studies is to develop statistical methodology that maximizes power to identify genes responsible for human disease susceptibility while minimizing the number of false positives. Complex diseases are expected to be influenced by many disease susceptibility loci. It is fully expected that multiple SNPs can be found associated with each disease of interest, and the results of most GWA studies have been consistent with this expectation. Single marker analysis, testing each SNP independently, can lead to biased estimation of effect sizes and significance. Such biases are small for the SNPs that are not associated with the disease, so the estimated significance, i.e. p-values, can be used to prioritize a set of most significant SNPs. But for the set of most significant SNPs which are usually associated with the disease, SMA overlooks the contributions of other loci while considering a certain locus, and thus can give undesirable rankings when prioritizing the SNPs in the set. Therefore, a reasonable statistical model for complex diseases should take into account multiple SNP effects that influence the disease additively or epistatically.

Bayesian mixture models simultaneously consider the influence of many SNPs, and simulation results show they are nearly always more powerful than single marker analysis. Dimension reduction of the models is achieved by using Bayesian variable selection, which has advantages over classical methods such as stepwise regression and LASSO [44]. In addition to the intensive computational burden, stepwise regression is a greedy algorithm and often gets stuck at local optima. The choices of thresholds for keeping and deleting variables in the model are also arbitrary. In contrast, the proposed mixture model approach se-

lects variables in a Bayesian framework, and the posterior distributions of SNP effects provide more information for making decisions of variable selection. The stochastic characteristic of the approach helps it to avoid getting stuck in local optima at the cost of inflated run times.

The primary challenge of Bayesian mixture models is that they rely on MCMC algorithms for posterior estimation, which are computationally intensive. For example, it would take a desktop workstation months to apply the mixture model we present directly to a genome-wide study that involves a large number of SNPs (e.g. ~500K). Fortunately, because of the fact that most of the SNPs have negligible effects, acknowledging this can greatly accelerate the analysis. One can first prioritize a set of SNPs and then apply the mixture models to the selected set. Possible prioritizing methods include sorting the p-values from SMA, or using penalized-likelihood methods (e.g. LASSO) for screening SNPs. These fast methods can work relatively well for ruling out most SNPs with negligible effects, and leave thousands or hundreds of SNPs in the model. The proposed score-averaging method also provide an option to accelerate the dimension reduction. It considers the influence of SNP effects simultaneously and also takes averages among many different models, which seems a more reasonable method for further acceleration. In our simulations, the Bayesian mixture models with the score-averaging acceleration applied on the SNPs prioritized by p-values of SMA ($p < 0.1$) took 4-6 hours on a desktop workstation for human chromosome 1, and 2-4 days for a genome-wide study. If using a SMA screening of $p < 0.05$, one can run the methods at the scale of 100K SNPs within an hour and obtain reasonably good results.

The proposed Bayesian mixture models assume that each SNP effect follows

a 3-component mixture prior (2.5), while other Bayesian variable selection methods have proposed different forms of mixture priors: a) George and McCulloch [47] propose a two-component normal mixture prior for their SSVS method; b) Mitchell and Beauchamp [59] use a "spike and slab" prior which is a mixture of a uniform distribution and a point mass at 0; c) Genkin *et al*. [60] choose a "double-exponential" Laplace prior. These mixture priors all can be used for screening variables, but the 3-component mixture prior (2.5) is more appropriate for variable selection at a finer scale. It combines the advantage of a), b) and c), by explicitly modeling the probability of being exactly zero, and assuming normal distributions for nonzero effects. The symmetric distribution of c) imposes the same degree of shrinkage on SNP effects, while the different truncated normal distributions for positive and negative effects in (2.5) result in different degrees of shrinkage for positive and negative effects, and hence can better fit the data.

Another advantage of the proposed mixture models is that generalized linear models are considered to model the contribution of multiple risk factors to susceptibility of diseases simultaneously. The linear combination part in (2.1) or (2.2) can be easily extended to include more terms such as gene-gene interactions, environmental effects, and gene-environment interactions. After modifications, the mixture models then can explicitly take epistasis into account. Simulation shows the performance of the logistic model (2.1) and that of the probit model (2.2) are comparable, and users can choose the model they desire for case-control studies.

The first-generation of genome-wide association studies have been analyzed with only the simplest of single-SNP tests, and given the investment of time

and money in these projects, optimization of the statistical methodology is particularly well motivated. To meet this challenge, we propose Bayesian mixture models and show in simulation that they are nearly always more powerful than single marker analysis, suggesting their potential capacity of identifying more disease susceptibility loci.

Implementation

A software termed SAGAS (Statistical Analysis for Genome-wide Association Studies) was written in C and is under active development. It has implemented the methods proposed in this chapter, including the logistic mixture model and probit mixture model, as well as the linear mixture model, which can accommodate either binary traits or continuous, quantitative traits. The software accepts the plain ASCII format as well as the binary PLINK format for input. The command line interface (CLI) mimics that of PLINK, and many parameters (e.g. MCMC-related) can be changed by the user. More features will be added to the software, and it is expected to eventually be made publicly available.

CHAPTER 3

AN EFFICIENT LINEAR MIXED MODEL THAT ACCOUNTS FOR POPULATION STRUCTURE

3.1 Introduction

Genome-wide association studies (GWA, or GWAS) use genetic variation across the entire genome for identifying associations with observations, or the presence or absence of a disease or condition. It has drawn tremendous interest and been proved to be successful and powerful not only in humans [26], but also in many other organisms including mice [79], dogs [3], cattle [80], *Arabidopsis thaliana* [27], etc. A major concern, however, is the elevated false positive rate [19, 35] often resulting from the confounding by population stratification and cryptic relatedness. Population stratification refers to individuals in the sample coming from different populations, and cryptic relatedness is the unknown genetic relationship between individuals [81]; these two combined is termed sample structure [42], or more generally, population structure (We use this terminology throughout the chapter for convenience, although different ones can be found in literature). Population structure exists in both natural populations [82, 39] and model organism experiments [83, 40, 41]. The confounding, if not carefully accounted for, can cause spurious associations that appear significant but actually are not due to quantitative trait loci (QTL) or nucleotides (QTN) [33, 34]. It is noted that the problem exists for other association studies besides GWAS, for example, in the studies of maize [39], where the fast linkage disequilibrium (LD) decay makes GWAS unavailable [84].

Although careful designs of studies may avoid population structure, they

may be unavailable and thus have limited effectiveness for GWA studies. Many researchers hence have used statistical methods to account for the confounding. Various methods have been proposed including genomic control (GC) [36], structured association (SA) [37], principal component analysis (PCA) [38], and linear mixed models (LMM) [39]. Genomic control, assuming a similar structural effect on all loci, uses genetic markers to estimate and rescale test statistics to account for the inflation caused by population structure. It does not change the rankings of p-values, making it less helpful for genome-wide scans [40]. More importantly, GC is less powerful in the presence of strong confounding, as might be seen in model organisms [38, 39]. Structured association uses genotype data to assign individuals into subpopulations, and conditions on these assignments (stored in the Q matrix as in STRUCTURE) for any subsequent analysis. SA usually needs input from investigators about the number of subpopulations, and cannot capture the cryptic relatedness in the sample. Also assuming a small number of ancestral populations [41] like SA, principal component analysis replaces Q in SA by principal components (PCs), which have been found effective in representing population structure by using a few major axes. For example, recent reports have shown that strong PCs are strongly correlated with longitudes and latitudes [40, 85]. It also avoids the heavy computation as needed in STRUCTURE. However, it is not obvious to decide how many strong PCs to be used in the subsequent analysis. Recently, a promising method relying on linear mixed models has drawn attention, especially in the past few years. The idea is to model population stratification in the fixed effects and model the relatedness in the random effects whose covariance matrices take the form of certain similarity matrix. Specifically, the fixed effects accounting for population stratification can use population assignments (the Q matrix) [37, 39] or princi-

pal components [38], taking the advantages of SA or PCA; the similarity matrix uses some IBD kinship or IBS matrix. Recent studies have shown that LMM can effectively account for population stratification and genetic relatedness simultaneously [39, 40, 86, 41].

A potential problem of using LMM is the computational burden. Many implementations suffer substantial burdens in computation when applied to GWA studies with thousands of individuals genotyped at hundreds of thousands of loci, even though they consider one locus at a time. One part of the computation comes from estimating population assignments if the Q matrix from STRUCTURE is used. As mentioned before, this can be avoided by using principal components instead. But fitting the mixed model itself can still be computationally intensive. The method by Yu *et al*. [39] implemented in TASSEL takes hours to scan hundreds of SNPs on hundreds of strains. The efficient mixed model association (EMMA) proposes a single-dimensional optimization for estimating variance parameters, but is still slow for a large number of individuals and markers. To further speed up the use of LMM, Kang et al. [42] propose an expedited version of EMMA (EMMAX), while Zhang *et al*. develop a compression method as well as "population parameters previously determined" (P3D). These methods avoid recomputing variance parameters by imposing certain assumptions and are shown to work for GWA studies.

All the methods discussed so far usually scan all SNPs and test one at a time. However, many genes may underlie complex traits, for which testing one SNP at a time might not fully realize the potential of GWA studies. Some multilocus methods might be more appropriate instead, since a weaker effect may be more apparent after other causal effects are already accounted for [48]. When

a multi-locus method is considered, variable selection becomes necessary and also a challenging problem (as mentioned in Chapter 2). Stepwise procedures based on information criteria (e.g. AIC [54] and BIC [55]), penalized likelihood (e.g. LASSO [44]) and Bayesian methods (e.g. SSVS [47]) have been proposed for variable selection. Many of these methods have the potential to be extended to mixed-effects models, which is still an active research area. Metrics including a R^2 statistic [87] and a BIC-like criterion [88] have been proposed for selecting effects in mixed models. Fan and Li [89] and Ni *et al.* [90] propose penalized likelihood methods for semiparametric mixed models in longitudinal studies, which include the special parametric case. Kinney and Dunson [71] propose Bayesian mixed models that simultaneously select fixed and random effects, which can be reduced to the case where only fixed effects are selected. Some of the methods are based on optimizations and aim at obtaining likelihood maxima or posterior modes, while other methods use Markov chain Monte Carlo (MCMC) to simulate posterior distributions. All the methods, including the ones discussed in Chapter 2, can work for our problem in theory. For example, additional terms modeling population stratification and cryptic relatedness can be added to the mixture models in Chapter 2, and the entire model can be estimated in a fully Bayesian way.

Keeping in mind that computational efficiency is a top priority, I decided to avoid using MCMC, and instead focus on simplified assumptions and fast algorithms. Some methods propose fitting a mixed model first, and then using the residuals to construct linear models for single-SNP tests [91, 92, 42]. Additionally, least angle regression (LAR) is a useful and fast variable selection algorithm for linear regression that is less greedy than traditional forward selection methods [93]. By combining the ideas of these methods, I propose a multi-locus

LMM-based algorithm that is both efficient and takes account for population stratification and cryptic relatedness.

3.2 Methods

3.2.1 Linear mixed-effect model

The unified LMM proposed by Yu *et al*. [39] takes the form of (with some changes in notation)

$$\mathbf{y} = X\boldsymbol{\beta} + Q\boldsymbol{\nu} + S\boldsymbol{\alpha} + Z\mathbf{u} + \mathbf{e} \tag{3.1}$$

where $\boldsymbol{\beta}$ is a vector of SNP effects (including an intercept), $\boldsymbol{\nu}$ is a vector of population effects (e.g. PC's), $\boldsymbol{\alpha}$ is a vector of fixed effects other than SNP or population effects, $\mathbf{u}$ is a vector of genetic background effects, $\mathbf{e}$ is a vector of random errors, X, Q, S, Z are corresponding design matrices for the effects (the first column of X contains all 1s), and $\mathbf{y}$ is the vector of phenotypic values. The model (3.1) features several fixed effects and one genetic random effect that account for several levels of relatedness. The variances of the random effects are assumed to be $\mathrm{Var}(\mathbf{u}) = \sigma^2 K$ and $\mathrm{Var}(\mathbf{e}) = \sigma_e^2 I_n$, where K is an $n \times n$ similarity matrix defining the degree of genetic relatedness, I_n is an $n \times n$ identity matrix, σ^2 is the genetic variance, and σ_e^2 is the error variance. Let $\delta = \sigma_e^2/\sigma^2$, which is the ratio of the variance parameters, and it holds that

$$\mathrm{Var}\begin{pmatrix} \mathbf{u} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} K & 0 \\ 0 & \delta I_n \end{pmatrix} \sigma^2$$

For simplicity, we consider a model with the same notations that is expressed

as

$$\mathbf{y} = X\boldsymbol{\beta} + Z\mathbf{u} + \mathbf{e} \tag{3.2}$$

which is also used in EMMA [41]. If normality is assumed, $\mathbf{u} \sim N\left(0, \sigma^2 K\right)$ and $\mathbf{e} \sim N\left(0, \sigma_e^2 I_n\right)$, then,

$$\mathbf{y} \sim N\left(X\boldsymbol{\beta}, \sigma^2 ZKZ^T + \sigma_e^2 I_n\right)$$

The full log-likelihood and restricted log-likelihood functions can be formulated as (can also be found in the EMMA paper [41])

$$l_F\left(\mathbf{y}; \boldsymbol{\beta}, \sigma, \delta\right) = \frac{1}{2}\left[-n \log\left(2\pi\sigma^2\right) - \log|H| - \frac{1}{\sigma^2}\left(\mathbf{y} - X\boldsymbol{\beta}\right)^T H^{-1}\left(\mathbf{y} - X\boldsymbol{\beta}\right)\right] \tag{3.3}$$

$$l_R\left(\mathbf{y}; \sigma, \delta\right) = l_F\left(\mathbf{y}; \hat{\boldsymbol{\beta}}, \sigma, \delta\right) + \frac{1}{2}\left[q \log\left(2\pi\sigma^2\right) + \log\left|X^T X\right| - \log\left|X^T H^{-1} X\right|\right] \tag{3.4}$$

where $H = ZKZ^T + \delta I_n$ which is a function of δ, $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimate of $\boldsymbol{\beta}$, and q is the length of vector $\boldsymbol{\beta}$, i.e. the number of SNP effects plus one (for the intercept).

3.2.2 Fitting the reduced model

To keep the computation cost manageable for GWA studies, one would like to avoid using MCMC and iterative computations. Several methods have considered the reduced LMM, i.e. an LMM without SNP effects included, which is expressed as

$$\mathbf{y} = \mu\mathbf{1}_n + Z\mathbf{u} + \mathbf{e} \tag{3.5}$$

where μ is the overall phenotypic mean, and $\mathbf{1}_n$ denotes a vector with n 1s. This is a special case of model (3.2) by letting $X = \mathbf{1}_n$, $\boldsymbol{\beta} = \mu$, and $q = 1$. Some methods estimate the variance parameters only once from (3.5) and, keeping the values fixed, apply them globally to each SNP for single-SNP tests [42, 43],

by arguing that the approximation is feasible given the assumptions that the SNP effects and the other non-genetic effects are independent and that the effect of each SNP on the phenotype is negligible for the purpose of estimating variance parameters. Other methods [91, 92] optimize the reduced model (3.5) with SNP effects excluded and estimate the residuals. The residuals are then fitted as responses in linear models with SNP effects added in, which is equivalent to keeping the estimate of $\boldsymbol{u}$ fixed in all subsequent analyses. Although the former method is desirable, we decide to adopt the latter one so that least angle regression (LAR) [93] can be directly applied to the linear models.

Kang *et al*. [41] provide a one-dimensional optimization for solving δ that maximizes either (3.3) or (3.4). From now on, let us assume that δ takes the estimated value using EMMA [41] that maximizes the full log-likelihood, and H can be computed given δ.

According to Henderson [94], the prediction of $\mathbf{y} = X\boldsymbol{\beta} + Z\mathbf{u}$ is $\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}} + Z\hat{\mathbf{u}}$ where $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ are any solution to

$$\begin{cases} X^T X\boldsymbol{\beta} + XZ\mathbf{u} = X^T\mathbf{y} \\ Z^T X\boldsymbol{\beta} + \left(Z^T Z + \delta K^{-1}\right)\mathbf{u} = Z^T\mathbf{y} \end{cases} \tag{3.6}$$

and we can further obtain

$$\hat{\boldsymbol{\beta}} = \hat{\mu} = \left(X^T H^{-1} X\right)^{-1} X^T H^{-1}\mathbf{y} = \frac{\mathbf{1}^T H^{-1}\mathbf{y}}{\mathbf{1}^T H^{-1}\mathbf{1}}$$

according to Henderson *et al*. [95], and

$$\hat{\mathbf{u}} = KZ^T H^{-1}\left(\mathbf{y} - X\hat{\boldsymbol{\beta}}\right) = KZ^T H^{-1}\left(\mathbf{y} - \hat{\mu}\mathbf{1}\right)$$

according to Henderson [96]. We can now compute the residual vector $\boldsymbol{\eta}$ from model (3.5), which is

$$\boldsymbol{\eta} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - X\hat{\boldsymbol{\beta}} - Z\hat{\mathbf{u}} = \mathbf{y} - \hat{\mu}\mathbf{1}_n - Z\hat{\mathbf{u}}$$

The variance parameters can also be computed following Kang *et al*. [41] Given δ, the MLE of σ^2, $\hat{\sigma}_F^2$, and the restricted maximum likelihood (REML) estimate of σ^2, $\hat{\sigma}_R^2$, can be computed as

$$\hat{\sigma}_F^2 = \frac{R}{n}, \qquad \hat{\sigma}_R^2 = \frac{R}{n-q}$$

where

$$R = \left(\mathbf{y} - X\hat{\boldsymbol{\beta}}\right)^T H^{-1} \left(\mathbf{y} - X\hat{\boldsymbol{\beta}}\right) = (\mathbf{y} - \mu\mathbf{1}_n)^T H^{-1} (\mathbf{y} - \mu\mathbf{1}_n)$$

Note that $q = 1$ for the reduced model (3.5), so the difference between the MLE and REML is small for a large n.

3.2.3 Least angle regression

After the residual $\boldsymbol{\eta}$ is obtained from the reduced model (3.5) , it can be fitted as a response in linear models where explanatory variables are the SNP effects. We prefer a multi-locus model that has a modest-sized set of SNPs included, and a fast algorithm should be available to find such a model. That is to say, there is a trade-off of prediction, model size, and computational efficiency. To achieve this goal, we choose to use LASSO [44] for variable selection. The LASSO searches for β , for a given t, in the optimization problem of

$$\begin{aligned} &\text{minimize} \quad \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 \\ &\text{subject to} \quad \|\boldsymbol{\beta}\|_1 \leq t \end{aligned} \tag{3.7}$$

For different t, there will be a LASSO solution for the model, i.e. a $\boldsymbol{\beta}$ with some elements possibly being zero. Since computational efficiency is regarded as a top priority, we adopt least angle regression (LARS) [93] to find the LASSO solutions.

LARS was proposed by Efron *et al*. [93] with an original goal to explain the similarities between models produced by the LASSO and Forward Stagewise algorithms [97]. It is similar to the classic forward stepwise procedure, but less greedy. Let a null model represent a linear model with only an intercept included. Efron *et al*. [93] describe the unmodified LARS with comparisons to the forward stepwise procedure. Both start with the null model,

> "and find the predictor most correlated with the response, say x_{j_1}. We take the largest step possible in the direction of this predictor until some other predictor, say x_{j_2}, has as much correlation with the current residual. At this point LARS parts company with Forward Selection. Instead of continuing along x_{j_1}, LARS proceeds in a direction equiangular between the two predictors until a third variable x_{j_3} earns its way into the 'most correlated' set. LARS then proceeds equiangularly between x_{j_1}, x_{j_2} and x_{j_3}, that is, along the 'least angle direction,' until a fourth variable enters, and so on" (pp. 411, [93]).

Like forward stepwise, LARS keeps adding new variables into the model without dropping any one.

Efron *et al*. [93] also propose a simple modifications of the LARS procedure, which is actually employed in our algorithm. The modification, providing an efficient computation of all LASSO solutions, is that, with certain conditions, the ongoing LARS step may remove a variable from the calculation of the next equiangular direction (details can be found in Efron *et al*.[93]). They show that this procedure gives all LASSO solutions. This way of solving LASSO is very appealing, since the computation is only at the same order of ordinary least squares (OLS) to calculate the full set of LARS models. We choose to use the

"lars" R package that implements the modified procedure.

3.2.4 Proposed algorithm

Now we present the algorithm for LMM that is efficient and involves multiple SNPs. The procedure is as follows:

1. Initially, the reduced model , $\mathbf{y} = \mu\mathbf{1}_n + Z\mathbf{u} + \mathbf{e}$, is fitted. The residual of this model is $\boldsymbol{\eta} = \mathbf{y} - \hat{\mathbf{y}}$, where $\hat{\mathbf{y}}$ is the fitted value vector.
2. Then the linear model $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}'$ is fitted and the variables are selected using least angle regression (LARS, [93]), where $\mathbf{e}' \sim N\left(\mathbf{0}, \sigma^2_{e'} I\right)$, $\sigma^2_{e'}$ is an unknown parameter.
3. The process is expected to stop when no more variables are selected by LARS. However, an early stopping can be considered using some criterion.

We evaluate the performance of the proposed algorithm through simulations.

3.2.5 Choice of the relatedness matrix

A key component in the linear mixed model is the correlation matrix K, by which cryptic relatedness is actually accounted for. In the analysis of pedigrees, variance component analysis can use a K estimated from the known pedigree structure [98, 99]. For population-based GWA studies, one can estimate relatedness using a large number of SNPs. Loiselle *et al*. [100] estimate K to approximate IBD by adjusting the probability of IBS between two individuals with the

average probability of IBS between random individuals. This method has been implemented in SPAGeDi [101], and used by Yu *et al*. [39] and Zhang *et al*. [43]. However, it was found in simulations and also noted in literature [42] that the IBD kinship coefficients do not guarantee the K matrix be positive semi-definite, making it difficult for the correlation matrix. Some perturbation can be added to the matrix to make it positive semi-definite, but it is not a statistically sound solution. Other choices of K include IBS matrices and the Balding-Nichols (BN) kinship matrix [102]. IBS matrices can be guaranteed to be semi-definite positive [41], and be estimated using PLINK [78] or EMMA [41]. A haplotype-based or a simple IBS matrix has been found empirically more robust than the IBD kinship matrix[40, 41]. Moreover, a recent study [42] took either the simple IBS or the Balding-Nichols (BN) kinship matrix as the surrogate of sample structure and showed the two methods have a very high concordance to each other. Given the aforementioned comparison, we decided to use the simple IBS matrix as the K matrix for our algorithm. Since comparing the performance of different K's is not the focus of this study, in our simulations the phenotypic variance is simulated based on the K we calculated from the data.

The phenotypic variance is $\sigma^2 K + \sigma_e^2 I$, indicating that each individual phenotype has the variance $\sigma^2 + \sigma_e^2$, and that the ith and jth ($i \neq j$) phenotypes have the covariance $\sigma^2 K_{i,j}$. A smaller $\delta = \sigma_e^2/\sigma^2$ suggests that most of the variance is due to random error instead of genetic background. It is observed that the IBS matrix (e.g. the one estimated by PLINK) hardly has elements that are exactly zero. Many of the off-diagonal elements will be less than 1, and may be around 0.5-0.7. In the case that individuals are unrelated, the estimate of σ^2 will be much smaller than that of σ_e^2

3.2.6 Information criterion

The LARS procedure computes from the null model until a saturated model is reached, giving all the LASSO solutions. However, it could be preferable to early stop when the model size is small. An information criterion [88] similar to Bayesian Information Criterion is used for this purpose which, denoted IC, is expressed as

$$IC = -2l_F(\mathbf{y}; \boldsymbol{\beta}, \sigma, \delta) + df \cdot \log(n) \tag{3.8}$$

where $l_F(\mathbf{y}; \boldsymbol{\beta}, \sigma, \delta)$ is the full log-likelihood, df is the degree of freedom (usually the number of SNPs in the model plus 1 for the intercept and plus 1 for the random effect), and n is the sample size. We propose an early stop at the point where IC stops decreasing and begins increasing.

3.3 Simulations and Application

3.3.1 Simulation setup

Simulations are based on data sets with genetic relatedness in the sample. Given limited availability of such data sets to us, we choose the Utah residents with Northern and Western European ancestry from the CEPH collection (CEU) in HapMap [9] as the sample in our simulations. Individuals with a high missing rate (>10%), SNPs with a high missing rate (>10%) or a low minor allele frequency (<5%) are all removed. SNPs on sex chromosomes are also removed for simplicity. After filtering the Phase 3 data, there are 165 individuals left in the sample, each with genotypes of 252,302 SNPs on autosomes. A pairwise IBS

matrix is computed using PLINK and shown in Figure 3.1.

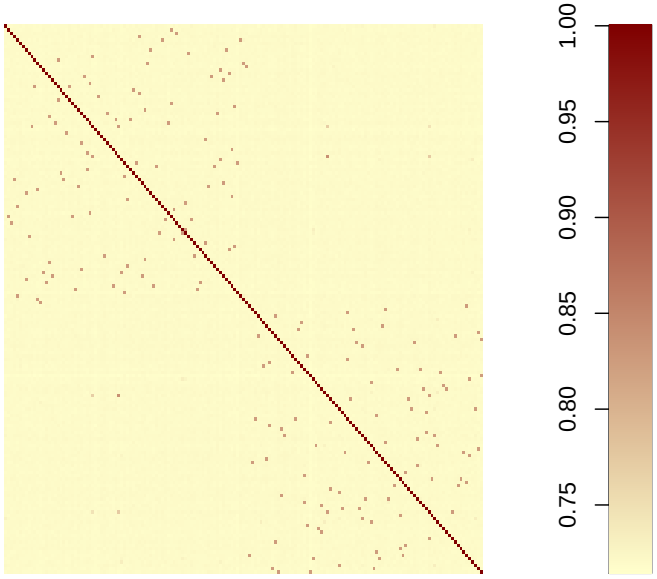

Figure 3.1: The IBS similarity matrix for the simulated data computed using PLINK.

To simulate the QTLs and phenotypes, we followed the similar procedures with Yu *et al*. [39] and Kang *et al*. [41]. First, we simulated phenotypes from a multivariate normal distribution. The variance is called background variance, Σ_b, which includes the variance of the genetic background excluding the SNP effects and also includes the variance of a random error vector. That is to say,

$$\mathbf{y} \sim N(\mathbf{0}_n, \Sigma_b)$$

where

$$\Sigma_b = \frac{(n-1)}{tr(S_0 Z K Z^T S_0)} h_g^2 K + \left(1 - h_g^2\right) I_n = h_g^2 K + \left(1 - h_g^2\right) I_n$$

n is the sample size, $S_0 = I_n - \mathbf{1}_n \mathbf{1}_n^T / n$, and we assume a special case $Z = I_n$ for simplicity [41]. After the phenotypic vector is simulated, the sample background variance, σ_b^2, is computed.

With respect to σ_b^2, the percentage of variance explained by a QTL, π, can be

estimated as [39]

$$\pi = \frac{2p(1-p)\tilde{\beta}^2}{2p(1-p)\tilde{\beta}^2 + 1 - \frac{1}{n}} \approx \frac{1}{1 + \frac{1}{2p(1-p)\tilde{\beta}^2}} \tag{3.9}$$

where n is the sample size, p is the causal allele frequency of the QTL, $\tilde{\beta}\sigma_b$ is the effect of the QTL, and genotypes are assumed to be encoded in 0, 1, 2 (the count of the causal allele).

Now several fixed effects based on a set of randomly chosen causal SNPs, i.e. quantitative trait loci (QTLs), are added to the simulated phenotypes to form new phenotypes, and for each QTL, $\tilde{\beta}$ is chosen to achieve certain desired percentage of variance explained by this QTL relative to σ_b^2, an approximation of the background variance.

To evaluate the performance of our method, we considered three scenarios and apply three methods, our proposed algorithm, single maker analysis, and a naive LARS, to the simulated data sets. Single marker analysis is performed using PLINK. The naive LARS is to directly apply LARS to the phenotypes without consider a mixed model (i.e. not accounting for K). The details of the three scenarios are as follows:

Scenario 1: Assume $h_g^2 = 2/3$. 10 SNPs are randomly chosen as QTLs, each with $\tilde{\beta} \sim$ Uniform $[0.4, 0.6]$. 50 such replicates are considered.

Scenario 2: Assume $h_g^2 = 2/3$. 5 SNPs are randomly chosen as QTLs, each with $\tilde{\beta} \sim$ Uniform $[0.6, 0.8]$. 50 such replicates are considered.

Scenario 3: Assume $h_g^2 = 2/3$. 5 SNPs are randomly chosen as QTLs, each with $\tilde{\beta} \sim$ Uniform $[1.0, 2.0]$. 100 such replicates are considered.

Analyzing each of the data sets using our proposed algorithm requires less

than 1 minutes on a MacBook Pro (2.2GHz CPU, 2GB MAM), while applying EMMA to the same data set would take around 15 minutes.

3.3.2 Performance evaluation

It is straightforward for biologists to take a ranking method on the loci and pick the most significant or highly ranked ones for further investigation. When evaluating performance, one can look at the rankings of the loci and see how the causal loci are ranked in simulations. Ranks also have connections with false positives: suppose the SNPs are ranked by two methods, respectively. If the fifth causal SNP is ranked 10 by one method and ranked 20 by the other, then in order to see this causal SNP (and obtain 5 true positives), there will be 5 false positives for the first method and 15 false positives for the second one. Therefore, we may use the ranks as a measure of performance and compare different methods.

The ranks given by SMA are based on the p-values of the SNPs. For our proposed algorithm and the naive algorithm, both based on LARS, there are several ways to rank the SNPs. One is to use the order of the SNPs entering the model in the LARS procedure, and another option is to order SNPs by the value of the coefficients. The latter has some difficulty, since it depends on what model is chosen to look at the coefficients and how to choose such a model is not quite clear. After doing some simulations and comparisons, we decide to use the former one, i.e. the order of entering the model, as the way of ranking SNPs. Since LARS returns a model with degrees of freedom no more than the sample size n, the ranks given by the two LARS-based algorithms (the proposed

one and the naive one) cannot be greater than $n - 1$. Any SNPs not ranked will be regarded as negatives by these two algorithms.

For scenario 1, there are 500 QTLs in total in the simulations, and the 10 QTLs in each model together explain $40 \pm 5\%$ of the overall variance. Using single-SNP tests with Bonferroni's correction, 24 QTLs can be identified. Our proposed algorithm gives ranks to 49 QTLs, and the naive LARS gives 67 QTLs, the intersection of which contains 44 QTLs. Figure 3.2 shows the results of scenario 1. From panel b), we can see the largest percentage of overall variance explained by a single QTL is around 10%, suggesting each QTL has a small effect on the phenotype. Also the signs of rank differences do not seem to be correlated with the percentage of explained variance. Panels c) and d) suggest that, while both LARS-based algorithms give higher ranks than SMA, our algorithm gives higher ranks than the naive LARS algorithm, which may result from accounting for genetic relatedness.

For scenario 2, there are 250 QTLs in total in the simulations, and the 5 QTLs in each model together explain $44 \pm 4\%$ of the overall variance. Using single-SNP tests with Bonferroni's correction, 76 QTLs can be identified. Our proposed algorithm gives ranks to 133 QTLs, and the naive LARS gives 147 QTLs, the intersection of which contains 128 QTLs. The two LARS-based algorithms now give ranks to relatively similar number of QTLs. Figure 3.3 shows the results of scenario 2. From panel b), we can see the largest percentage of overall variance explained by a single QTL is around 20%, suggesting each QTL has a larger effect on the phenotype. The signs of rank differences still do not seem to be correlated with the percentage of explained variance. Both LARS-based algorithms give higher ranks than SMA. Our algorithm seems more likely to give

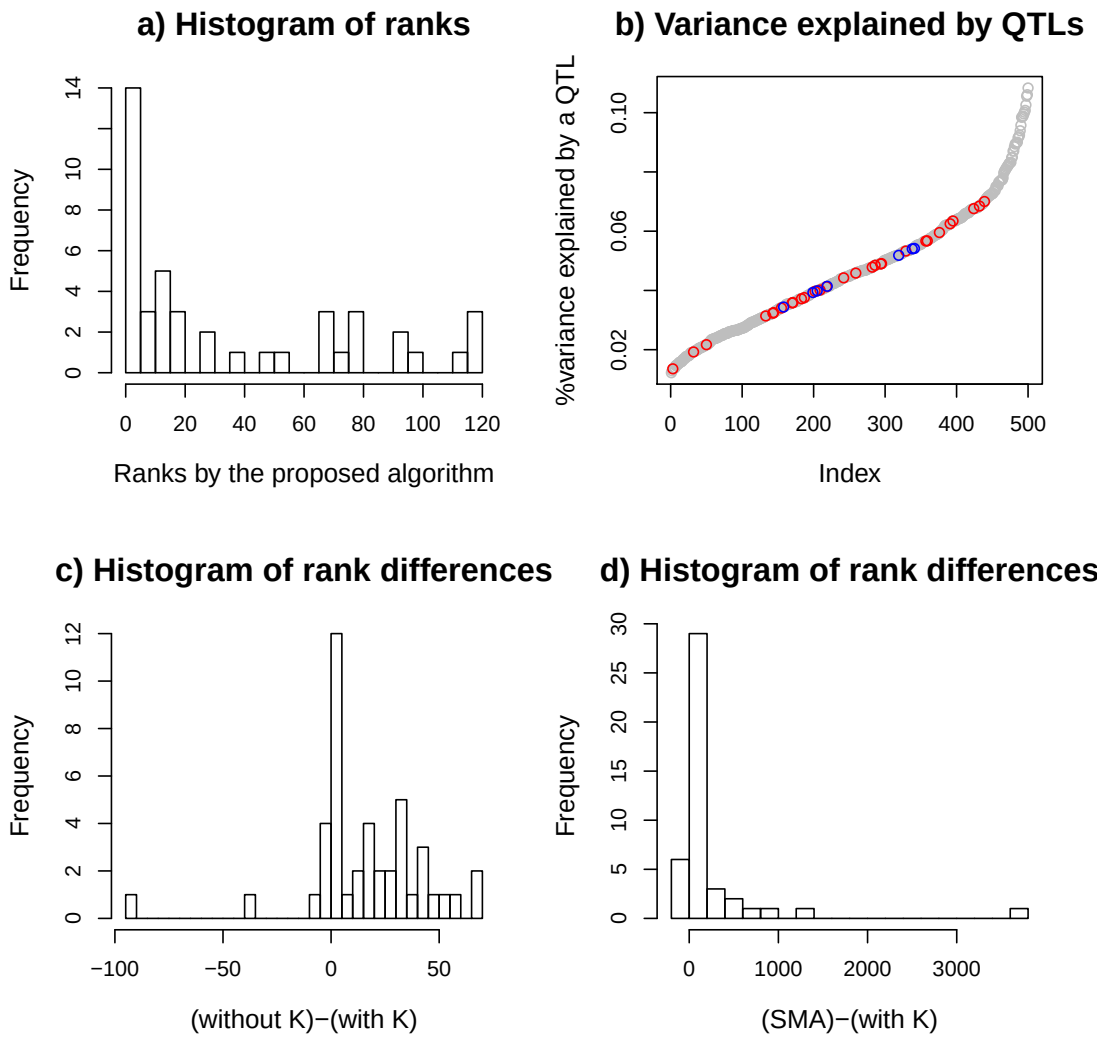

Figure 3.2: Simulation results of scenario 1. a) the histogram of ranks on true QTLs given by the proposed algorithm. b) the percentage of variance explained by each QTL across all the replicates. The red circles indicate QTLs for which the proposed algorithm gives higher ranks than the naive LARS, and the blue ones indicate QTLs for which the proposed algorithm gives lower ranks. Note that the largest percentage of variance explained by a QTL is around 10 %. c) the histogram of rank differences between the proposed algorithm and the naive LARS. A positive difference indicates that the proposed algorithm gives higher ranks. d) the histogram of rank differences between the proposed algorithm and single marker analysis. A positive difference indicates that the proposed algorithm gives higher ranks.

higher ranks than the naive LARS algorithm, and sometimes the differences can be large, suggesting that the ranks of QTLs might be elevated due to accounting for genetic relatedness.

The last scenario simulated 500 QTLs in total, and in each model, the 5 QTLs together explain $76 \pm 3\%$ of the overall variance, suggesting a higher heritability than the former two scenarios. Also each QTL seems to have a very large effect on the phenotype, with the maximum percentage of overall variance explained around 40%. All methods perform well for these large-effect SNPs: single-SNP tests with Bonferroni's correction can find 499 QTLs, our proposed algorithm gives ranks to 464 QTLs, and the naive LARS gives 471 QTLs, the intersection of which contains 464 QTLs. Although all three methods give high ranks to the QTLs, from Figure 3.4 we can still observe that, our algorithm can give high ranks to some QTLs while the ranks by the other two are rather low.

From these scenarios, we may conclude that:

1. QTLs with large effects are less subject to the influence of population structure, but for those with small or modest effects, the influence can be apparent, and should be accounted for.
2. Multi-locus methods could possibly make better use of GWA studies for complex traits. By accounting for other causal effects, some weak effects may be more apparent.
3. Our proposed algorithm, although simple and straightforward, provides an efficient solution of accounting for relatedness for GWA studies.

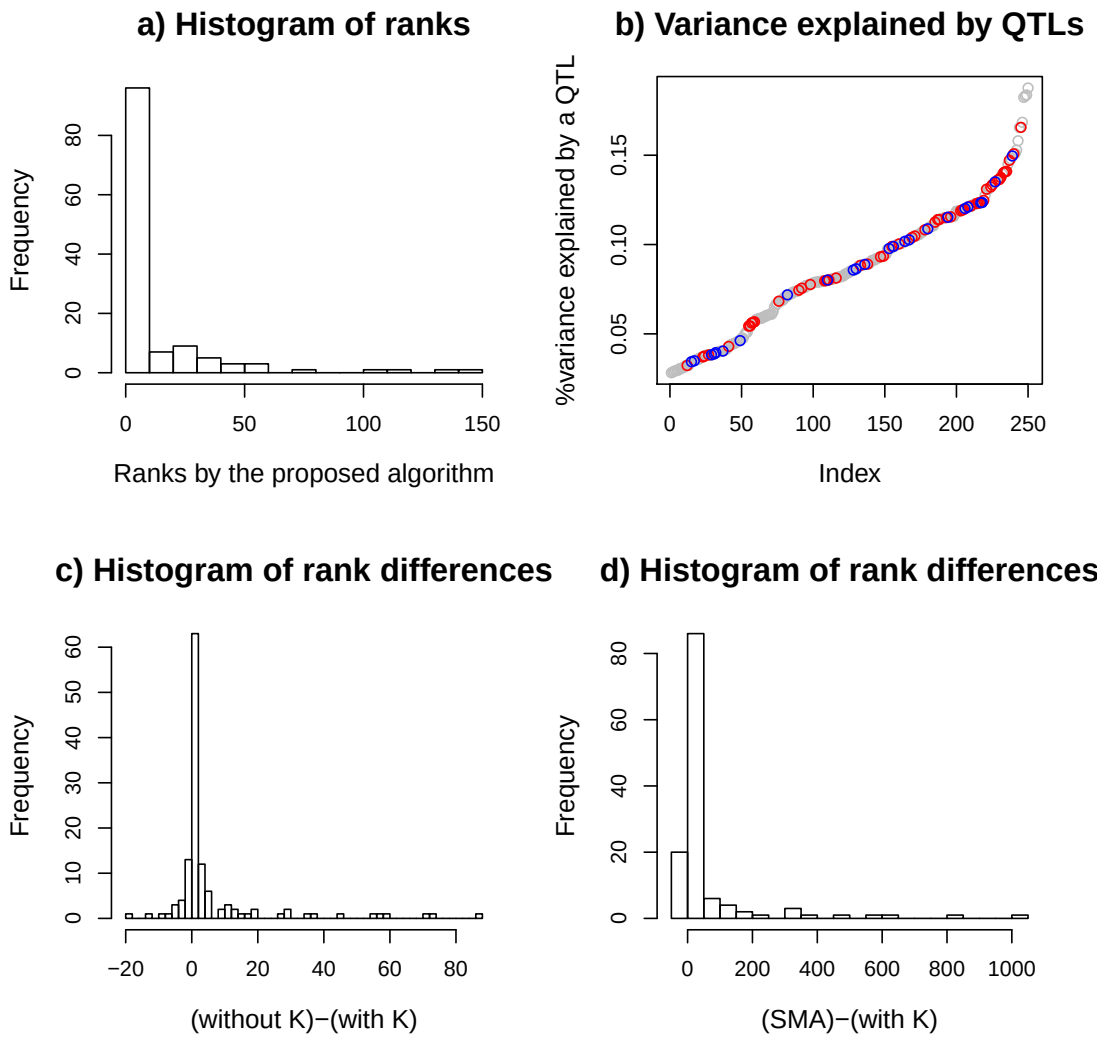

Figure 3.3: Simulation results of scenario 2. a) the histogram of ranks on true QTLs given by the proposed algorithm. b) the percentage of variance explained by each QTL across all the replicates. The red circles indicate QTLs for which the proposed algorithm gives higher ranks than the naive LARS, and the blue ones indicate QTLs for which the proposed algorithm gives lower ranks. Note that the largest percentage of variance explained by a QTL is around 20 %. c) the histogram of rank differences between the proposed algorithm and the naive LARS. A positive difference indicates that the proposed algorithm gives higher ranks. d) the histogram of rank differences between the proposed algorithm and single marker analysis. A positive difference indicates that the proposed algorithm gives higher ranks.

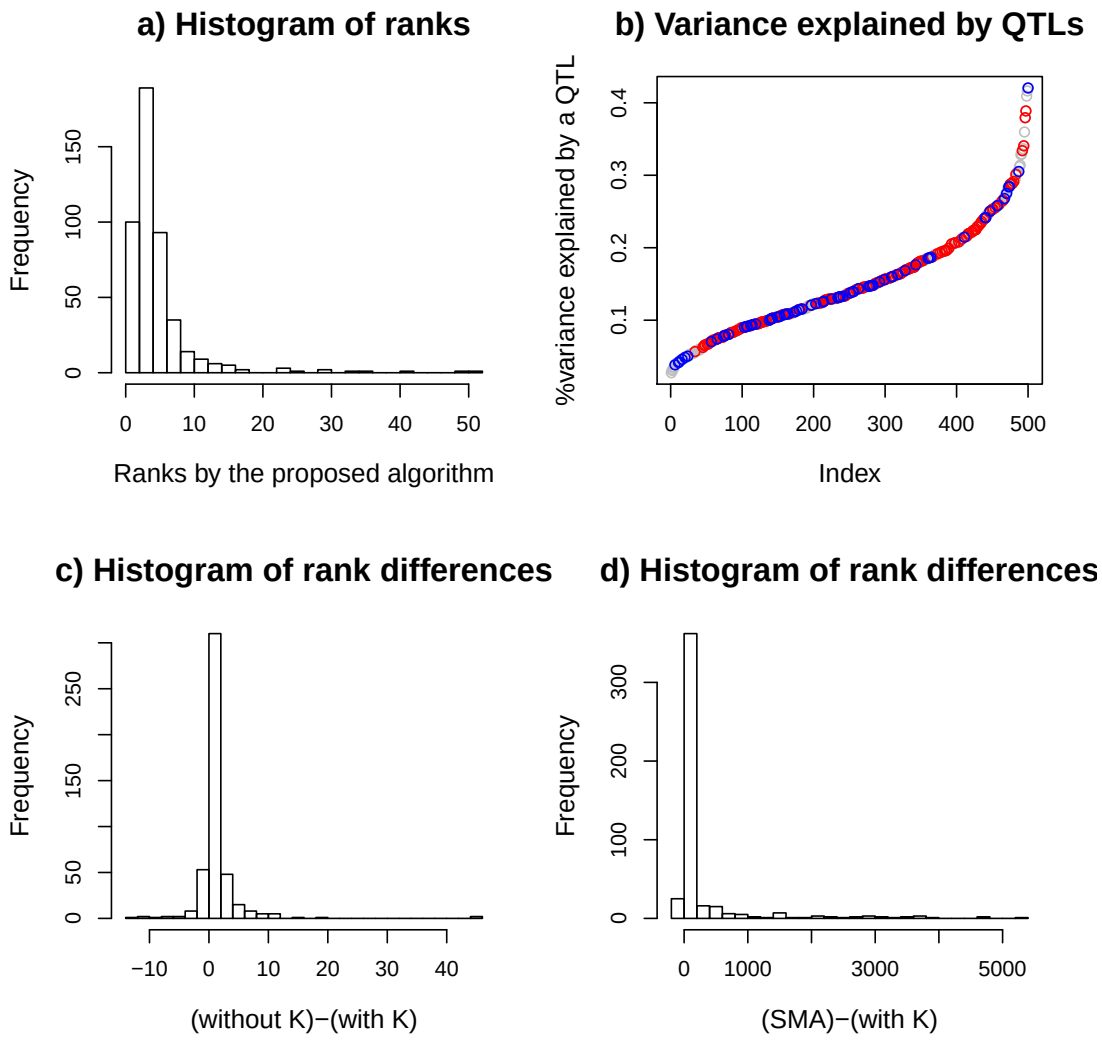

Figure 3.4: Simulation results of scenario 3. a) the histogram of ranks on true QTLs given by the proposed algorithm. b) the percentage of variance explained by each QTL across all the replicates. The red circles indicate QTLs for which the proposed algorithm gives higher ranks than the naive LARS, and the blue ones indicate QTLs for which the proposed algorithm gives lower ranks. Note that the largest percentage of variance explained by a QTL is around 40 %. c) the histogram of rank differences between the proposed algorithm and the naive LARS. A positive difference indicates that the proposed algorithm gives higher ranks. d) the histogram of rank differences between the proposed algorithm and single marker analysis. A positive difference indicates that the proposed algorithm gives higher ranks.

3.3.3 An application to a dog GWA study

We also applied the proposed algorithm to a GWA study on body weight in the domestic dogs. The study involves more than 1,000 dogs from ~80 breeds genotyped at about 60,000 SNPs. The average body weight of each breed is available and treated as the phenotypic value. Since the individual body weights are not available, the allele frequency of each SNP within breeds are computed and used to search for associations. After filtering, there are 79 breeds with breed average body weights and within-breed allele frequencies of 42,396 SNPs across the 38 autosomes and the X chromosome. To assess the genetic relatedness within breed, a breed-average IBS similarity matrix is used. (More details can be found in Chapter 4).

A linear mixed model is considered with breed average body weights (in logarithm) as the response, SNP allele frequencies as the fixed effects, and a random effect with the IBS matrix as its correlation matrix. Figure 3.5 shows the results of the proposed algorithm. The BIC-like information criterion suggests a 6-SNP model may be considered, and Figure3.6 plots the absolute coefficients of the first 6 SNPs entering the model in the LARS procedure. These SNPs actually have a very high concordance with the results we obtained using other methods (see Chapter 4). Given the small sample size and relatively small number of SNPs, the computing time is less than 1 minute, but running EMMA would take much more time on this data set.

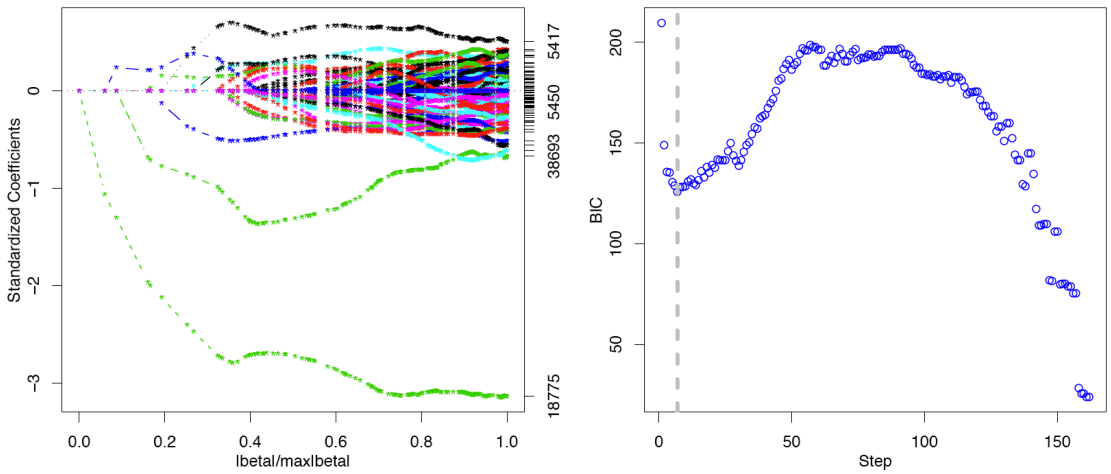

Figure 3.5: The results of the proposed algorithm applied to the dog body weight GWAS data. The left panel shows all the LASSO solutions computed using LARS, and the right panel shows the BIC-like criterion changes with LARS steps. The first local minimum of the IC is obtained at step 6, with 6 SNPs included in the model.

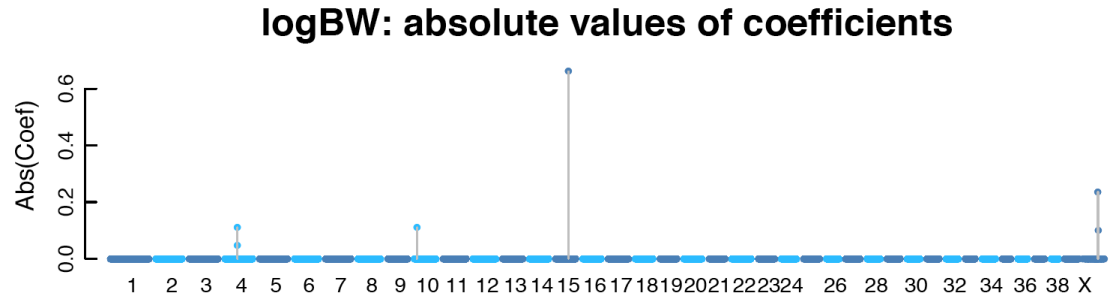

Figure 3.6: The absolute coefficients of SNP effects in the 6-SNP mixed model for body weight. The 6 SNPs are, in the order of entering the model, CFA 15:44226659, CFA X:106189665, CFA X:106866624, CFA 4:42392342, CFA 10:11440860, and CFA 4:42351982 (the 7th SNP is CFA 7:46842856).

3.4 Discussion

Population stratification and cryptic relatedness are inevitable problems for many GWA studies. Appearing evidence has shown that the linear mixed model method is a promising solution. However, fitting linear mixed models using classical methods for hundreds of thousands of SNPs, even one at a time, can be very computationally intensive and time-consuming, making it necessary to develop efficient algorithms. Moreover, to explore the most out of the potential of GWA studies, multi-marker methods may be considered instead of testing one SNP at a time independently. In this chapter, an efficient multi-marker linear mixed model algorithm is proposed to answer this need. In order to keep the computation cost modest and manageable, some assumptions and approximations have to be made, but some efficiency also comes from the use of efficient algorithms like LARS. To approach the problem, a linear model is constructed first by fitting a reduced mixed model and taking the residual as the new response, and LARS is run on the linear model to find LASSO solutions. Since we are interested in finding the SNPs that might be informative for causal loci or QTNs, we evaluate the performance of the proposed method and compare it to single-SNP tests by looking at the prioritized SNPs from the two methods. We have shown in the previous section that using multi-marker models can give higher ranks to causal SNPs, and accounting for relatedness can help elevate the ranks and thus reduce false positives.

However, it should be noted that a drawback exists for the proposed method, which is due to estimating the random effects only once and keeping them fixed. In our algorithm, the residual is computed and treated as the response of a linear model, so that LARS can be used directly. This is a strong assumption–

not only the variance parameters are kept fixed but also the random effects are unchanged once estimated and they are independent of the SNP effects. Aulchenko *et al*. [91] take a similar approach for single-SNP tests, and they remark that the method has a statistical power equivalent to that of the full optimization approach only for traits with low heritability. Although this is often the case in humans where many diseases have relatively low heritability, one may still want to relax this assumption. One possible solution is to modify LARS so that after each step the random effect can be reestimated based on the current set of explanatory variables (fixed effects), and use the new residual as the response of the linear model. However, whether it will improve the performance is still unknown. Moreover, extensive simulations are still necessary to evaluate the proposed method and its variants under more scenarios. An application to a dog GWA study has been considered here. This seems to be the only proper data set available to us right now, and it should be noted that this study may not be an ideal application for illustrating the power of the proposed method, as the genetic relatedness between pairs of breeds does not matter much for the phenotype [3] (although one needs to account for the confounding after all). A very recent study on *Arabidopsis thaliana* [27] may be considered for applications. All these will be covered in future research.

Although there still exist problems within the proposed method, it provides an algorithm to perform the requested tasks efficiently. Moreover, it also has some space and possibilities for improvement. Besides what have been discussed above, prediction is not the focus of this research, but accuracy of prediction (e.g. disease risk prediction) using multiple SNPs may be studied. In the case that the sample size is large, one may also combine the compression method by Zhang *et al*. [43] with ours to further speed up the analysis. We

agree with Efron *et al*. in their rejoinder to the discussions on LARS [103] that LARS-type programs are a good first step toward a solution, but hopefully not the last step, and suggest that our proposed algorithm should also be a good first step, but by no means the last one.

CHAPTER 4

GENOME-WIDE ASSOCIATION STUDIES ON COMPLEX MORPHOLOGICAL TRAITS IN DOMESTIC DOGS

4.1 Introduction

The dog was domesticated from the wolf, which may date back to 15,000 years ago or even more [14]. Humans have intentionally bred dogs for many generations and selected specific traits which are desirable for working or as a companion pet. Many breeds have been formed since the Victorian Era, and nowadays, there are more than 300 breeds all over the world, over 160 of which are American Kennel Club (AKC) [104] registered breeds. Although being the same species, *Canis familiaris*, these breeds represent a vast diversity, making the domestic dog the most diverse land mammal on the earth. Intensively selected for morphology and behavior, the dog can vary dramatically on its body weight, height, skull shape, and leg length, etc., with variation larger among breeds than within breeds [105, 106, 107]. The level of genetic diversity is also considerably less within any single breed than for all breeds combined [52]. The interbreed heterogeneity and intra-breed homogeneity have drawn the interests of scientists in investigating the genetic basis of complex traits in the domestic dog, especially morphological ones [1, 2, 108, 53].

Many examples of the traits that have underlying genetic variants identified in the past three years include skeletal size [2], coat color [109, 110], leg length [111], hairlessness [112], wrinkled skin [113], hair length, curl and texture [114], and presence of a dorsal fur ridge [115]. In particular, body weight is an important trait with variation much greater than that of humans. For example,

an adult Great Dane can have a weight 50 times that of an adult Chihuahua. Different from humans, where only loci with modest effects having been found and explaining a small proportion of variance [116], a single IGF1 allele is found to be a major determinant of small size in dogs [2]. It was not known though how additional loci may contribute to such variation should there exist any. A possible genetic model is that a few genes with large effects underlie most of these traits in dogs, whereas an alternative one is that a trait may be controlled by hundreds of genes each with very modest effects. It was our great interests to assess whether the majority of phenotypic variation among breed-affiliated dogs is a consequence of causal loci with larger effects, and also the extent to which domestication and artificial selection have shaped the dog genome [3].

To better understand the genetic basis of complex traits in the domestic dog, a high-density map of common genetic variation (the "CanMap" Project) has recently been developed [3]. The resource consists of more than 1,000 individuals from 80 AKC breeds as well as wild canids and Egyptian shelter dogs [117], each with more than 120,000 single nucleotide polymorphisms (SNPs) genotyped. Relying on the CanMap data set and phenotypic information available to us, genome-wide association (GWA) studies on more than 100 morphological traits of domestic dogs have been carried out. A group of researchers, including Carlos Bustamante, Adam Boyko, and me, have worked on a paper describing research and results related to the data set, and my responsibility was to carry out statistical analysis in search for associations. In this chapter I will present mainly the statistical analysis methods I used as well as the corresponding results in the study.

4.2 Results

4.2.1 Decay of linkage disequilibrium and distributions of long runs of homozygosity

Pairwise SNP linkage disequilibrium (LD) and runs of homozygosity (ROHs) greater than 1Mb for each individual were quantified using the genotype data from the 59 breeds with $\geq$ 10 individuals and a population of village dogs and wolves. LD extends over 1 Mb within every breed, and decays extremely rapidly across all dogs combined (Figure B.1), suggesting few IBD segments are shared across multiple breeds. Individuals from nearly every domestic dog breed have 10-50 ROHs greater than 10 Mb, which are both longer and more numerous than those of village dogs and wolves (Figure B.2). These observations suggest that human-directed breeding has reshaped the dog genome in a radical way such that it leaves less genetic and phenotypic variation within breeds. [3]

The calculations of these results on LD decay and ROH were done by coauthors in [3], and I was responsible for generating Figures B.1 and B.2. These results are included here for convenience and completeness, as they are related to the interpretation of the association results.

4.2.2 Initial study of body weight

The initial data set of CanMap v2.0 contains 61,468 SNPs and 1,659 individuals. The number of duplicated samples is 151. The total number of breeds in the

data set is 85 and the number of domestic breeds is 80 .

For IGF1, there are 143 breeds with known allele frequencies for the IGF1 allele in Sutter *et al.* [2]. Figure 4.1 plots the IGF1 allele frequencies versus the breed average body weight of these 143 breeds using the data in [2], and there seems to be a strong association between IGF1 and breed average of body weight.

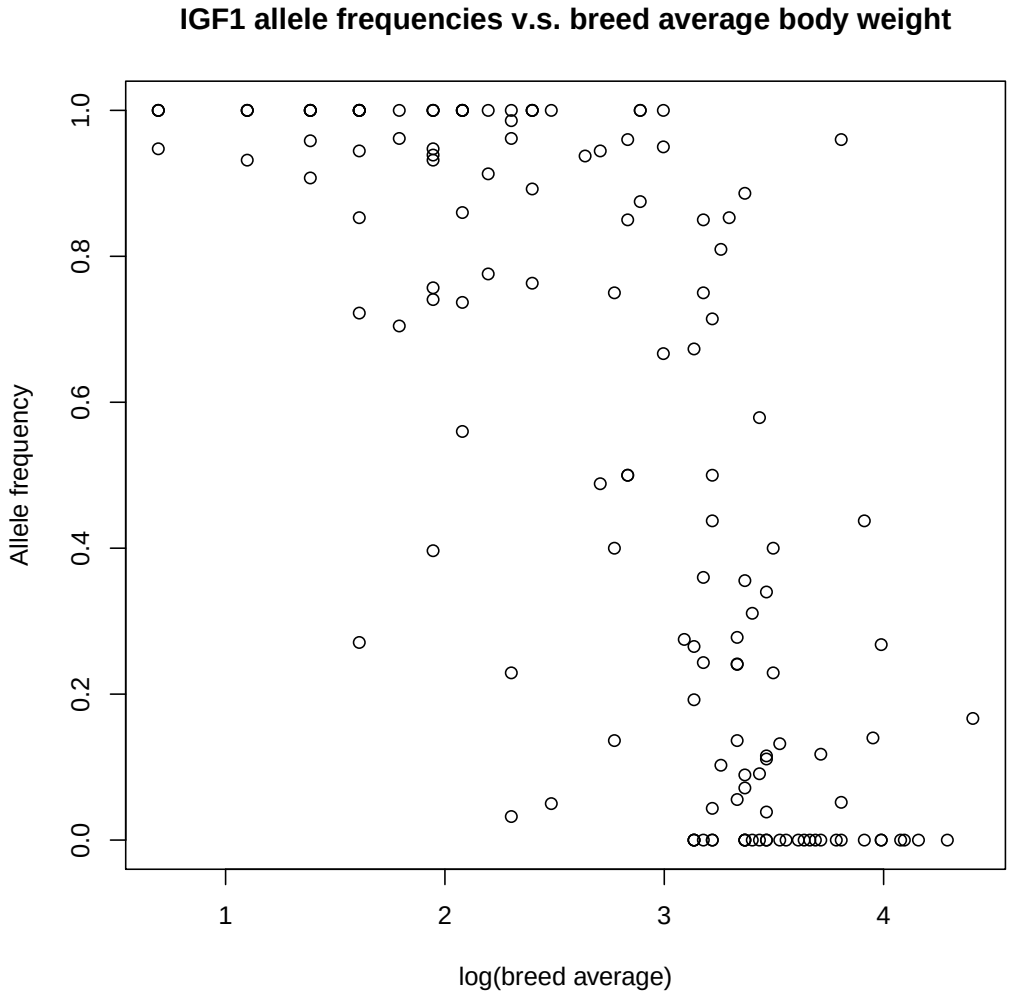

Figure 4.1: Association between IGF1 allele frequencies and breed averages of body weight for 143 domestic dog breeds using data in [2]

As an initial study, I extracted 844 dogs from 76 breeds and used their breed averages of body weight as phenotypes. By assigning the average body weight of each breed to all the dogs in that breed, I used the "pseudo-" phenotypes by overlooking the within-breed variation. Then a linear mixture model, which is similar to Zhang *et al*. [63], was used to search for associations of multiple SNPs simultaneously.

Two models are considered, the dog-genotype model (4.1) and the breed-frequency model (4.2),

$$\log(WT_i) = \mu_{\text{geno}} + \sum_{j=1}^{M} X_{ij}\beta_j, \quad i = 1, \cdots, N \tag{4.1}$$

$$\log\left(WT^{(k)}\right) = \mu_{\text{freq}} + \sum_{j=1}^{M} f_{ij}\gamma_j, \quad k = 1, \cdots, K \tag{4.2}$$

where WT_i, $(i = 1, \cdots, N)$ are the "pseudo-" phenotypes (body weights) for all N dogs, $WT^{(k)}$, $k = 1, \cdots, K$ are the breed averages for all K breeds, μ_{geno} and μ_{freq} are the intercepts for the two models, X contains the SNP genotype information with β_j modeling corresponding effects, and f contains the allele frequency information with γ_j modeling corresponding effects.

Both the dog-genotype model and the breed-frequency model are used for analyzing either autosomal SNPs only or all available SNPs. The dog-genotype model gives several strong signals which seem biologically related to body weight (Tables 4.1 and 4.2), while the breed-frequency model only yields one strong signal, IGF1, (Tables 4.3 and 4.4) which is in concordance with what has been identified before [2].

For the strong signals from the dog-genotype model, each was searched for genes and their biological functions using the UCSC genome browser. Het-

Table 4.1: Top 10 hits from the dog-genotype model regressed on autosomal SNPs using a linear mixture model. The posterior probabilities are estimated using a Gibbs sampler.

Rank	Posterior	SNP	Nearby gene
1	1.0000	CFA 15.44226659	IGF1
2	0.9998	CFA 10.11440860	HMGA2
3	0.9989	CFA 4.42351982	STC2
4	0.9973	CFA 18.23298242	CD36
5	0.8849	CFA 3.93851186	THOC4
6	0.5656	CFA 26.16250034	PLAC2, FLJ42957
7	0.4219	CFA 10.17108695	
8	0.2868	CFA 28.38460604	ADAM12
9	0.2765	CFA 23.38650936	
10	0.2754	CFA 6.63106915	

Table 4.2: Top 10 hits from the dog-genotype model regressed on all SNPs.

Rank	Posterior	SNP	Nearby gene
1	1.0000	CFA 15.44226659	IGF1
2	1.0000	CFA 4.42351982	STC2
3	0.9447	CFA 10.11440860	HMGA2
4	09240	CFA X.105988061	
5	0.7910	CFA 3.93851186	THOC4
6	0.6420	CFA 18.23298242	CD36
7	0.4979	CFA 7.44130632	INSRR, NTRK1
8	0.4239	CFA X.88183292	
9	0.3355	CFA 26.16250034	PLAC2,FLJ42957
10	0.2617	CFA 11.30002381	

Table 4.3: Top 10 hits from the breed-frequency model regressed on autosomal SNPs

Rank	Posterior	SNP	Nearby gene
1	1.0000	CFA 15.44226659	IGF1
2	0.1955	CFA 26.27358275	DRG1
3	0.1928	CFA 7.46842856	SMAD2
4	0.1777	CFA 36.5905408	
5	0.1630	CFA 28.4504108	
6	0.1606	CFA 23.49638253	
7	0.1550	CFA 23.38650936	
8	0.1482	CFA 22.43697584	EEF1A1
9	0.1479	CFA 7.46837936	SMAD2
10	0.1439	CFA 27.8941962	OR6S1

Table 4.4: Top 10 hits from the breed-frequency model regressed on all SNPs

Rank	Posterior	SNP	Nearby gene
1	0.9965	CFA 15.44226659	IGF1
2	0.1612	CFA X.110689568	
3	0.1541	CFA X.110850946	
4	0.1535	CFA X.110355686	
5	0.1535	CFA X.110622104	
6	0.1514	CFA 37.29932072	
7	0.1462	CFA X.110743613	
8	0.1451	CFA X.110365285	
9	0.1444	CFA X.110370844	
10	0.1432	CFA 27.8941962	OR6S1

erozygosities, both observed and expected, were computed using PLINK for three categories of dogs:

small dogs body weight≤10,

medium-sized dogs 10<body weight≤30, and

large dogs body weight>30.

The small dog category contains 280 dogs of 26 breeds, the medium one contains 326 dogs of 30 breeds, and the large one contains 238 dogs of 20 breeds.

Hits that seem interesting are shown in Figure 4.2 with heterozygosity runs. Heterozygosity ratios between large and small dogs, and those between medium-sized and small dogs, both show elevated values in the neighborhood of the hits (true for both observed and expected values). Figure 4.3 shows the smoothed heterozygosity runs in the regions of the hits as well as the location of the genes nearby, and it seems more clear that there exist genes near the hits that exhibit varying heterozygosity among different categories of dogs. In particular, the hits implying candidate genes nearby (in parentheses) include CFA 15.44226659 (IGF1), CFA 10.11440860 (HMGA2), CFA 4.42351982 (STC2), CFA 18.23298242 (CD36), and CFA 28.38460604 (ADAM12), which are the top 1, 2, 3, 4, and 8 hits, respectively, from the dog-genotype model on autosomal SNPs. IGF1 mediates many of the growth-promotion effects of growth hormone (GH; MIM139250) and has been studied previously [2]; HMGA2 has connections to diet-induced obesity; STC2 is related to growth restriction, reduced bone and skeletal muscle growth, and organomegaly; CD36 may be functional for fatty acid transport and related to growth hormone-releasing peptide; and ADAM12 may be implicated in biological processes involving muscle development.

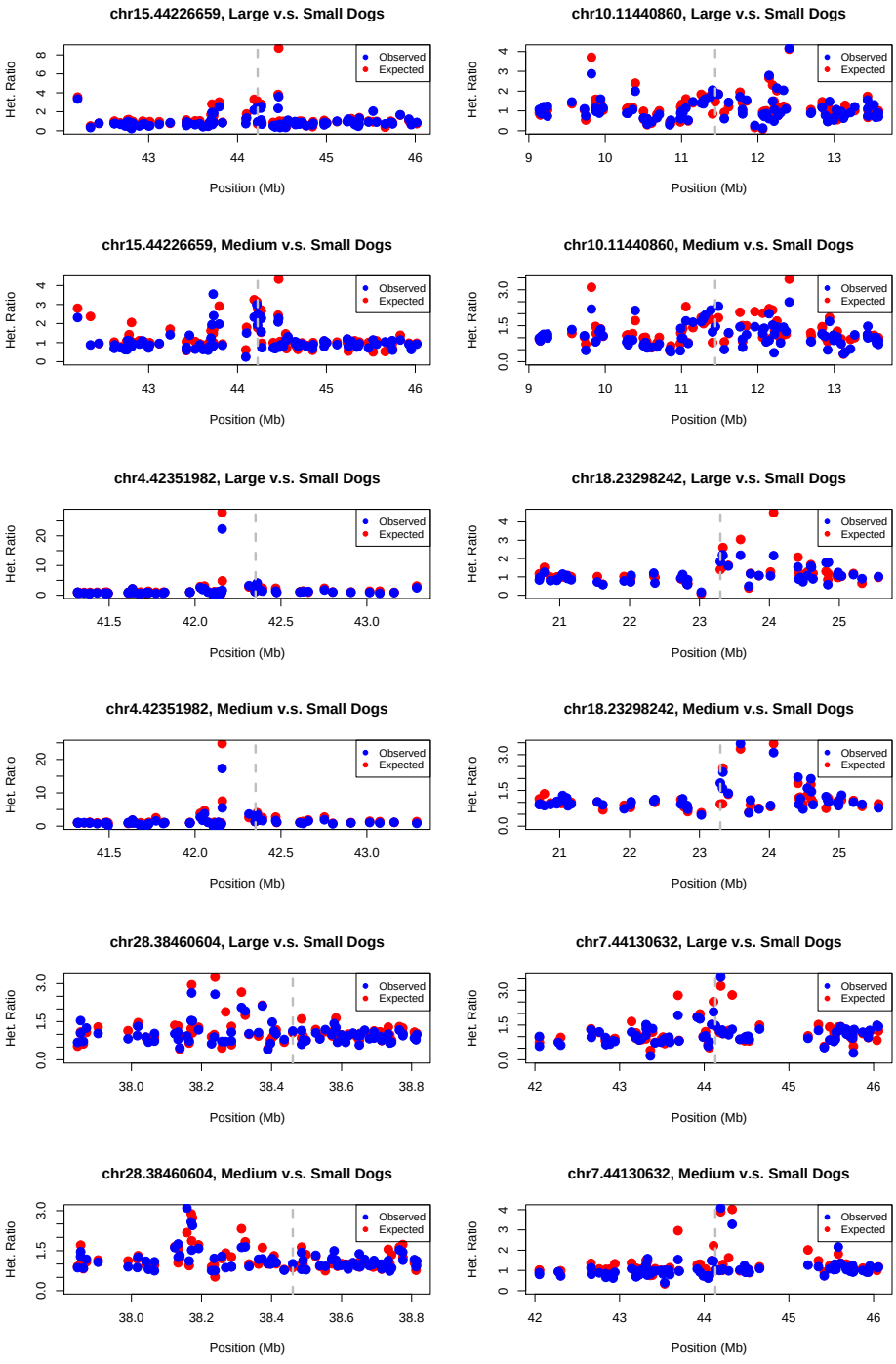

Figure 4.2: Heterozygosity runs in the neighborhood of the chosen SNPs. Heterozygosity ratios between large and small dogs as well as between medium-sized and small dogs are both plotted. Blue dots are observed values, and red dots are expected values. Calculations are done in PLINK.

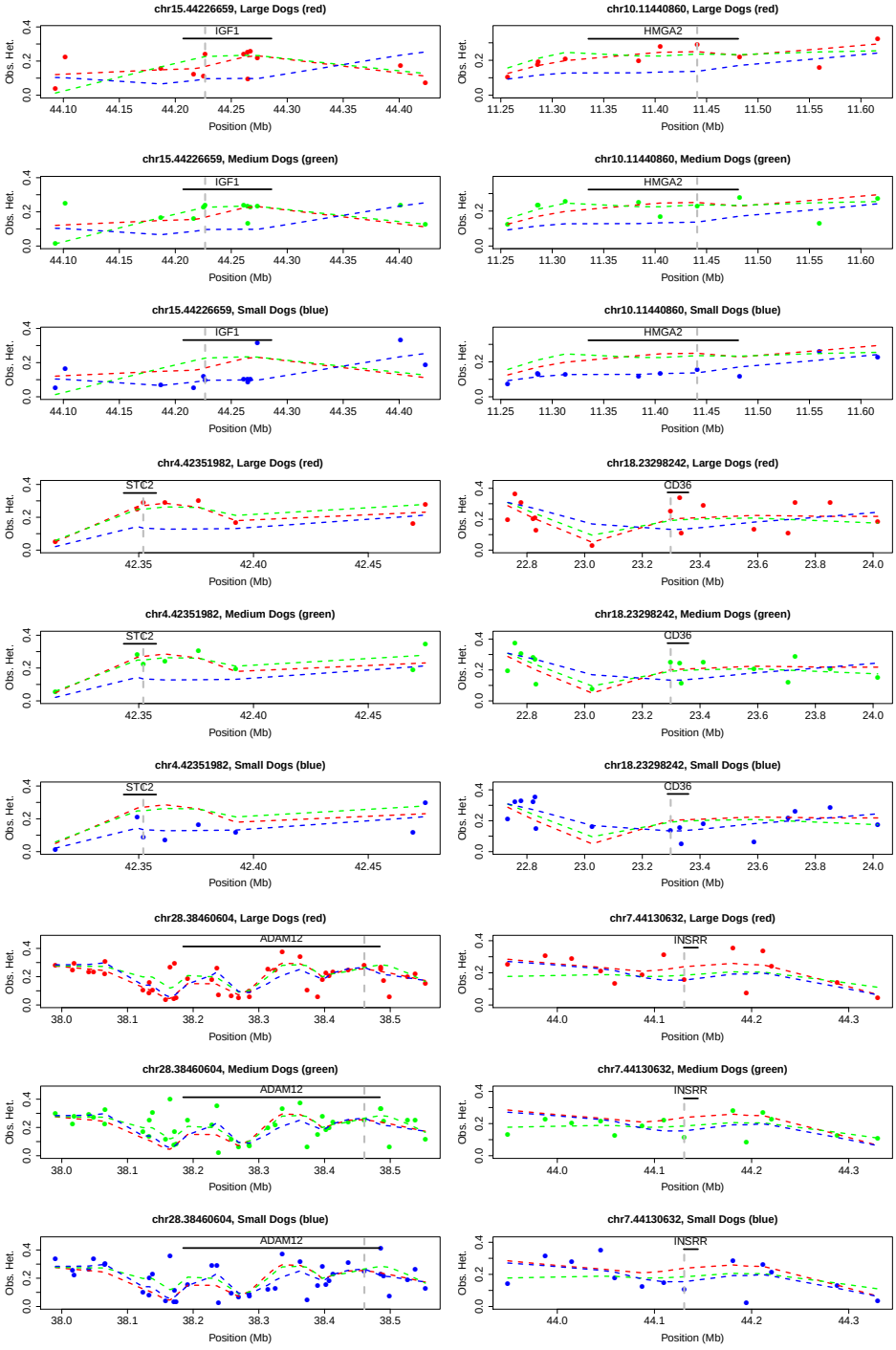

Figure 4.3: Smoothed heterozygosity runs using lowess. Red dots are observed heterozygosity for large dogs, green ones are for medium-sized dogs, and blues ones are for small dogs. The dash lines in corresponding colors and the smoothed lines. The black solid line indicates the gene region in the neighborhood, if there is any.

Figure 4.4 shows the genome-wide scan for body weight using single-SNP χ^2-tests as well as the Bayesian linear mixture model. Also shown are the model fitting using top 10 SNPs from the Bayesian scans as well as predictions on breed dogs and village dogs. It shows that the top hits are similar between the two different methods, and using a 10-SNP predictive model gives good predictions on both the breed and village dogs. There seems to be no significant biases between the predictions for males and and those for females. It is not clear, though, whether the 10-SNP model is the best one or not.

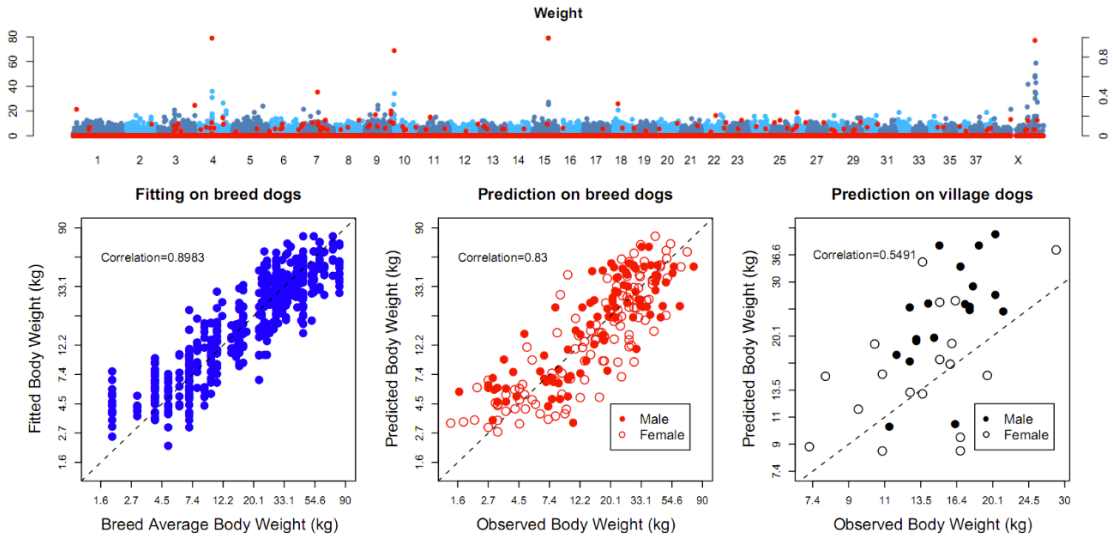

Figure 4.4: Genome-wide scan for body weight and model fitting. The upper panel shows the genome-wide scan results, with SMA p-values (in blue) and posteriors from Bayesian linear mixture models (in red). The lower panel contains three plots, showing the results of model fitting, and predictions on breed dogs as well as village dogs.

4.2.3 Initial study of external tape measurement traits

After the initial study on body weight, we decided to pursue GWA studies on other morphological traits. There are 1,035 dogs from 67 breeds with tape mea-

surements (Table 4.18). The ages of the dogs are also known, which were used to decide whether an individual is adult (more than one year old, in our study) or not. Whether the ears and tail have been cropped or not is also available.

Breed averages and standard deviations are estimated from the tape measurements. The traits of interest include height at withers, height at tail base, head width, ear length, etc. For those traits (e.g. ear length) that have measurements of two sides, the averages between the two sides were used as the values for the traits. Obviously some traits are highly correlated.

Among the 1,035 dogs, 563 dogs of 58 breeds have been genotyped in the CanMap project. Since the breed averages are used as pseudo-phenotypes for individual dogs in each breed, all dogs genotyped in the CanMap project, not restricted to the 563 dogs, are considered as long as the breed averages are available for their corresponding breeds. The genotypes and phenotypes can then be used to search for associations.

There are around 200 dogs in the CanMap project with genotypes as well as individual tape measurements available. This set of data can then be considered a validation set for a fitted model.

Single marker analysis without/with principal components as covariates

The principal components derived from the genotypes were used as covariates in the linear mixture model to account for possible population structure. The first 5 principal components, PC's, (explaining the largest proportion of variance) were included in the model. Single marker analyses both without and with the covariates were carried out. SNP CFA 15.44226659 (IGF1) is the top

hit for most measurements, and the results without and with PC's are similar, except for head width, where IGF1 appears more significant after PC's are included.

Bayesian variable selection for analyzing the 21 size measurements

For the external traits of tape measurements, the linear mixture model has the largest posterior probabilities (around 1.00) on CFA 15.44226659 (IGF1) for almost all the traits, with the only exception for head width, where CFA X.104724717 is the top hit with posterior probability of 1.00. Other hits that are shared by many traits are CFA 18.23298242, CFA 10.11440860, and CFA 4.42351982.

Take height at withers for example, the top hit is CFA 15.44226659. A 10-SNP predictive model is fitted, and then used to make predictions on the validation set. Figure 4.5 shows the genome-wide scan results, the scatter plots showing the correlations between the fitted (predicted) and breed averages (observed values), as well as adjusted R^2 when new terms (SNPs) are added into the model up to the 10-SNP model is obtained. The hit on chromosome 18 is CFA 18.23298242, which is in concordance with previous finding [111].

Another trait to note is the outside ear length, shown in Figure 4.6 with the genome-wide scans and model fitting. Besides the hit on chromosome 15 which is IGF1, there are several other hits that looks interesting and may be true, for example the hits on chromosome 10 and chromosome 3. Again, a 10-SNP model seems fitting the data well, although an over-fitting is possible. The changes in adjusted R^2 suggest that less SNPs may be necessary.

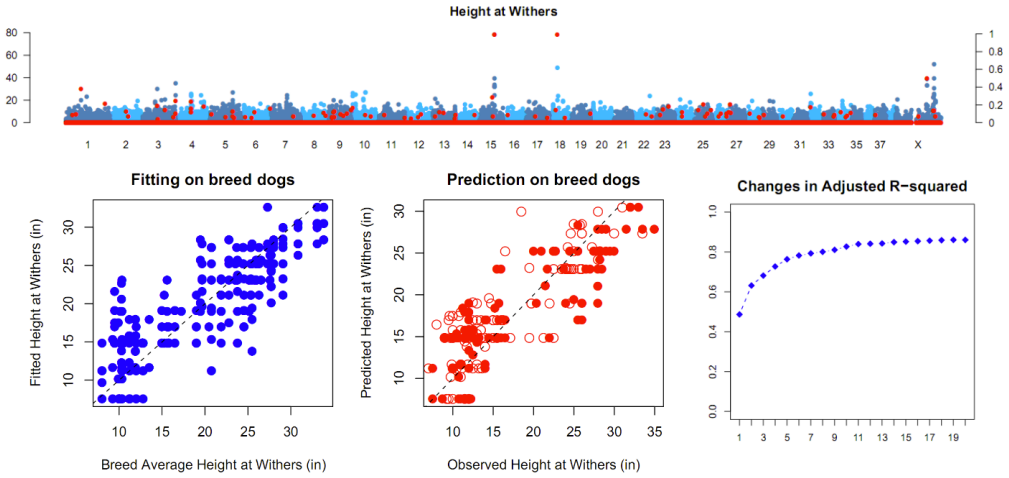

Figure 4.5: Genome-wide scan for height at withers and model fitting. The upper panel shows the genome-wide scan results, with SMA p-values (in blue) and posteriors from Bayesian linear mixture models (in red). The lower panel contains three plots, showing the results of model fitting, and predictions on breed dogs as well as the changes in adjusted R^2 when SNPs are added into the model sequentially.

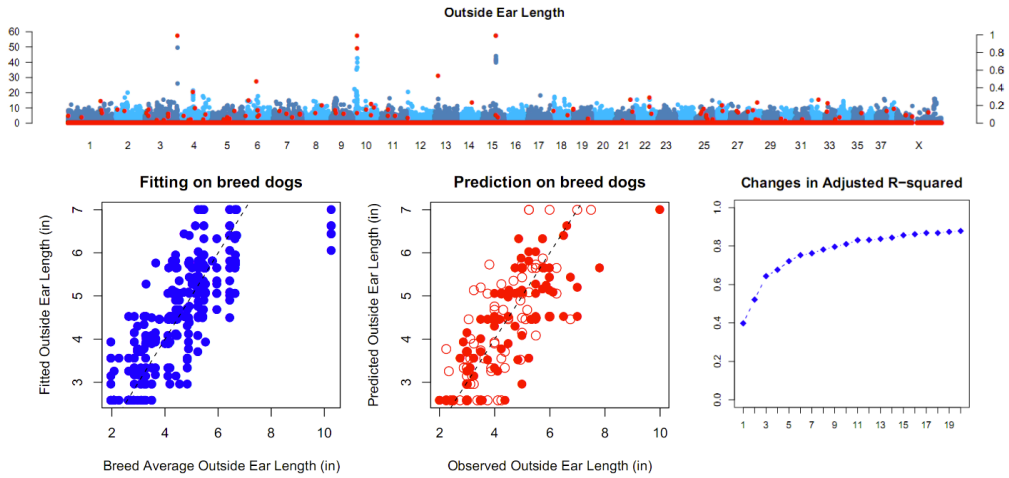

Figure 4.6: Genome-wide scan for ear length and model fitting. Genome-wide scan for height at withers and model fitting. The upper panel shows the genome-wide scan results, with SMA p-values (in blue) and posteriors from Bayesian linear mixture models (in red). The lower panel contains three plots, showing the results of model fitting, and predictions on breed dogs as well as the changes in adjusted R^2 when SNPs are added into the model sequentially.

As an attempt to see how significant the top hits are, Bayesian linear mixture models were also run on SNPs with the top 10 hits excluded. Figure 4.7 shows the changes in correlations for height at withers: the left panel shows the results of the original scan, and the right panel shows the results of the scan with top 10 hits excluded. From the figure, we can see that the top 10 hits in the original scan have contributions in predictions. A 2-SNP model using the top 2 hits can obtain a correlation of around 0.8, while about 10 SNPs are needed to get a similar correlation after the top 10 hits are excluded. It suggests that the top 2 hits are likely to be true, which are the IGF1 hit and the hit on chromosome 18 (in concordance with previous results).

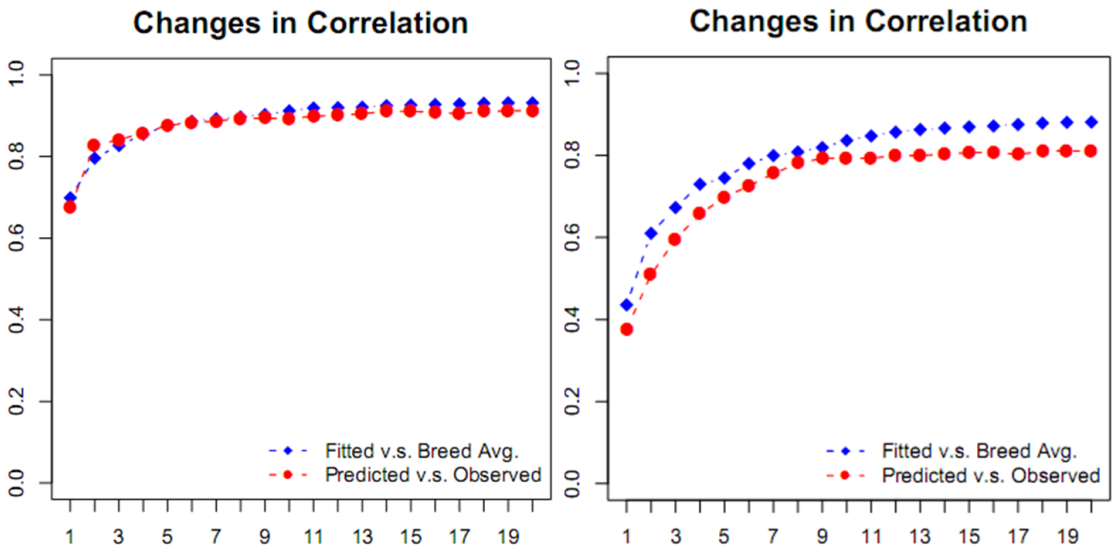

Figure 4.7: Changes in correlation before and after the top 10 SNPs are removed. The left panel shows the changes in correlation when the top 1 to 20 SNPs are added into the model sequentially, and the right one shows the changes when the top 21 to 40 SNPs are added into the model sequentially.

Bayesian variable selection for analyzing the 10 ratios of size measurements

The linear mixture model has also been directly run on the ratio traits. It was found that CFA 15.44226659 as the top hit for head ratio, and snout ratio, while CFA .18.23298242 is a top hit for leg ratio, chest ratio, girth ratio, and foot ratio. The latter finding is interesting, as researchers have identified an expressed Fgf4 retrogene nearby is associated with breed-defining chondrodysplasia in domestic dogs [111].

Bayesian variable selection for analyzing the residuals of size measurements

The ratio of the size measurements takes the ratio between two measurements. We found that it would make more sense to calculate the residuals of the traits after regressing out body weights. For a trait Y, consider the residual $Y_{residual}$ from regressing Y on body weight WT, e.g. using the R code `lm(log(Y)~log(WT))$residual`. The continuous residual $Y_{residual}$ takes values in the real domain $(-\infty, \infty)$ after using the logarithm transformation on both Y and WT. Bayesian scans can then be applied to these residual traits. Figure 4.8 shows the scans of single marker analysis as well as the Bayesian mixture model on the residual traits of height at withers, lower hind leg length, and head length.

Regressed on body weight, the traits now on longer have CFA 15.44226659 as the top hit, given the SNP being the major determinant for body weight. CFA 18.23298242 is among the top hits for height at withers and lower hind leg length, which again has an interesting connection with previous finding [111]. For the residual of height at withers, the breeds that have large positive values

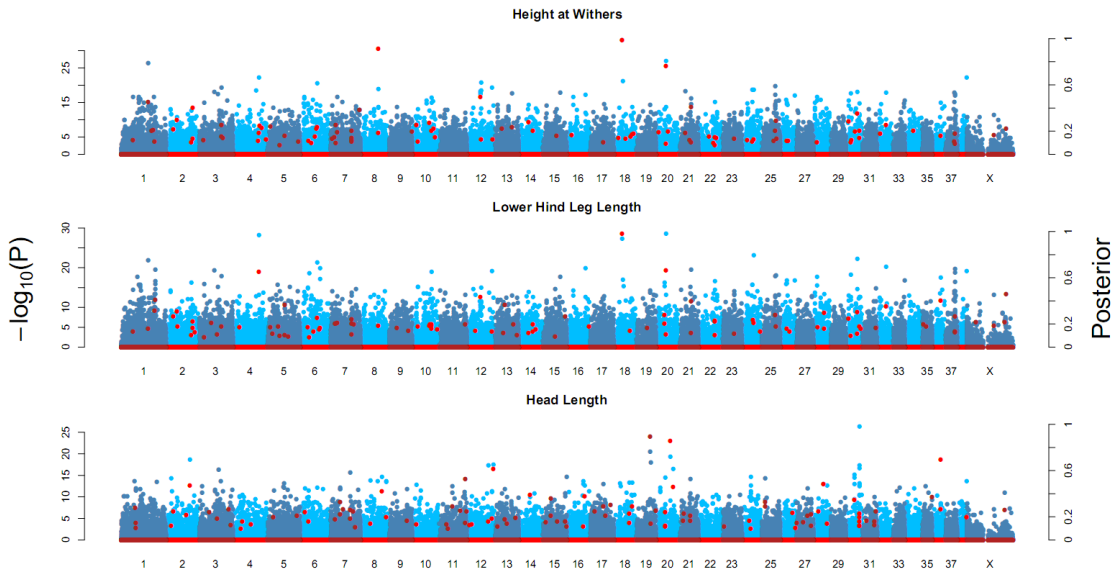

Figure 4.8: Genome-wide scans on residual traits. The blue dots show $-\log_{10}(p)$ from SMA, and the red dots show the posterior probabilities from Bayesian scans.

include Italian greyhound, whippet, Saluki, Afghan hound, Yorkshire terrier, Shetland sheepdog, greyhound, borzoi, Pomeranian, and Irish wolfhound; the breeds with small negative values include basset hound, dachshund, Scottish terrier, cardigan welsh corgi, bulldog, Sussex spaniel, and bull terrier. Figure 4.9 shows the allele frequencies of two hits, CFA 18.23298242 and CFA 20.26188392, for the residual trait of height at withers. The hits, especially CFA 18.23298242, are fixed in some breeds with extreme phenotypes, i.e. relatively tall or short.

It is not clear, however, whether the top hits for the residual trait of head length are legitimate given limited information that is available to us.

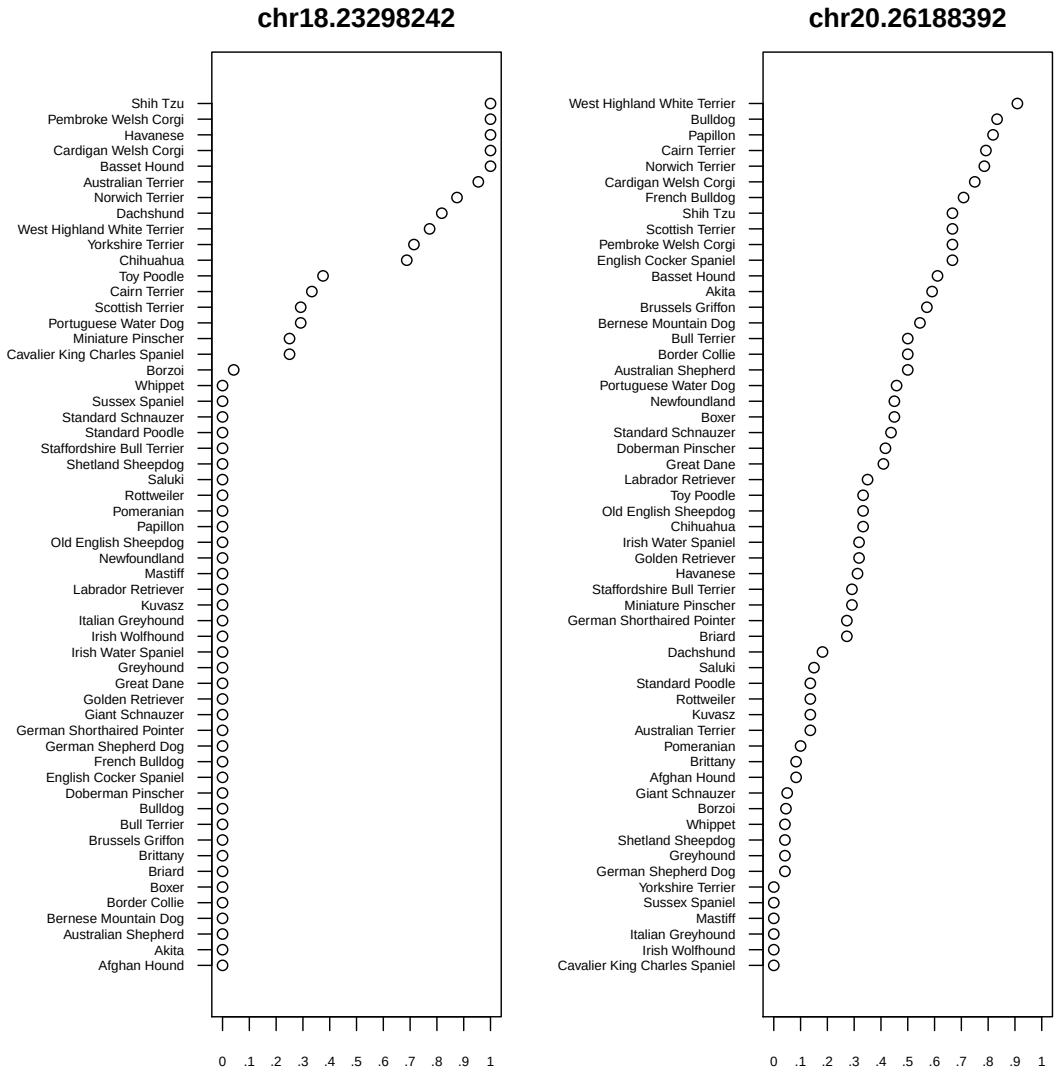

Figure 4.9: Allele frequencies of hits within different breeds for height at withers residuals

Bayesian variable selection for analyzing the first 5 principal components derived from the size measurements

The principal components of size measurements were also considered, and the first 5 principal components were used as quantitative traits. The linear mixture model has the top hit of CFA 15.44226659, CFA 7.45610981, CFA 18.23298242,

CFA 4.72681733, and CFA 1.96328985, respectively, for the 5 principal components. The 1st component is thought to represent the body size diversity among breeds, while the 3rd component is possibly related to limb morphology.

For the first component, the top hit for the trait is CFA 15.44226659. A 10-SNP predictive model is fitted, and then used to make predictions on the validation set, where the corresponding phenotypes are computed based on the previous loadings and individual measurements. Figure 4.10 shows the changes of several metrics when new terms are added into the model up to the 20-SNP model, and the scatter plots showing the correlations between the fitted (predicted) and breed averages (observed values).

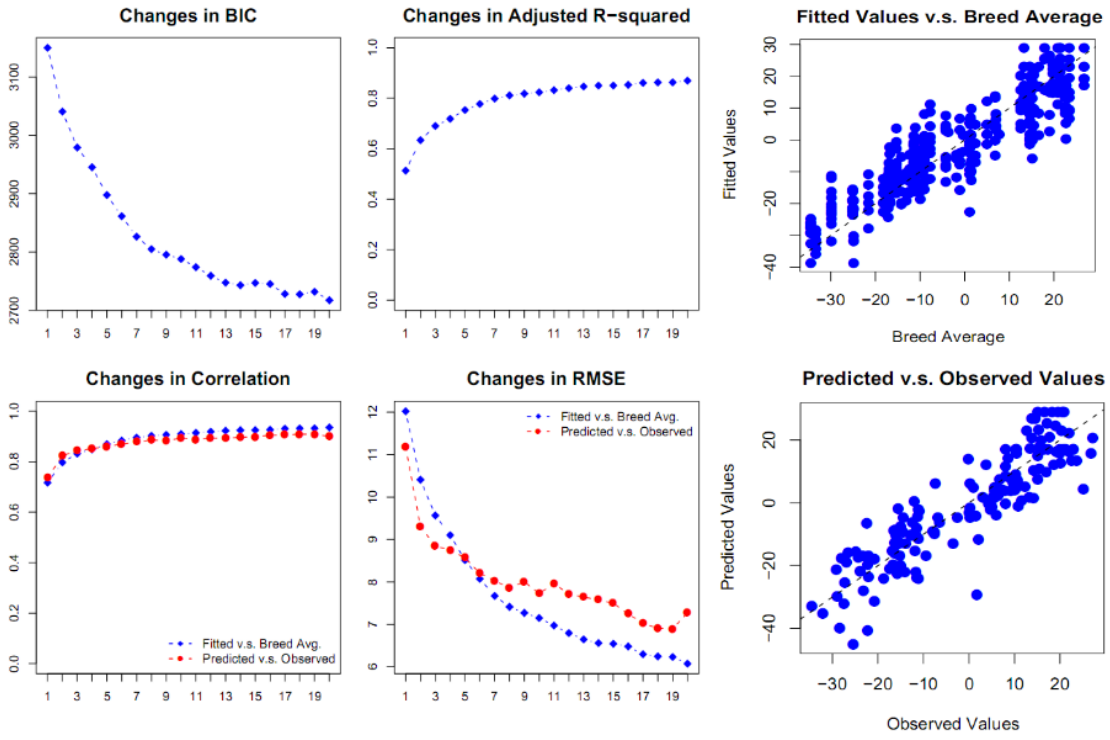

Figure 4.10: Bayesian scan hits for the first principal component of tape measurement. The four panels on the left show the changes of different metrics as top hits are added into the predictive model. The two panels on the right show the scatter plots for model fitting and predictions.

4.2.4 Initial study of ear floppiness

Ear forms vary among different breeds of domestic dogs: some dogs have floppy ears, some have erected ears, while others fall in between the two categories. On the other hand, variance in form is minimal within breeds. It would be interesting to study the genetic basis of ear floppiness as defined by breeds. We take the breeds with either floppy ears or erect ears and classify them into two groups, and use the group classification as a binary trait. Using the genotypes available, we carried out genome-wide single-SNP scans as well as Bayesian scans for the binary trait. Figure 4.11 shows the genome-wide scan for ear floppiness and also the model fitting. There are three top hits, not in strong linkage disequilibrium, in the region on chromosome 10 ranging 10,393,318-11,440,860bp. Using the three hits, the fitted model can obtain an accuracy over 90%, while the baseline is around 50%. The plot on the right in the lower panel (Figure 4.11) shows the fitted probabilities of the response equal to 1 (in red dots), and the grey areas indicate that a prediction is correct, i.e. the fitted probability$\geq$ 0.5 for a response that is truly 1. The results suggest that at least the region on chromosome 10 is worthwhile being investigated in follow up studies.

4.2.5 Permutation for significance of terms in multiple regression

Permutations are used to assess the significance of signals by using a threshold for posterior probabilities. The threshold is chosen for each trait such that the posterior probability p satisfies $P(p > t) \approx 0.10$. Table 4.5 shows the results for

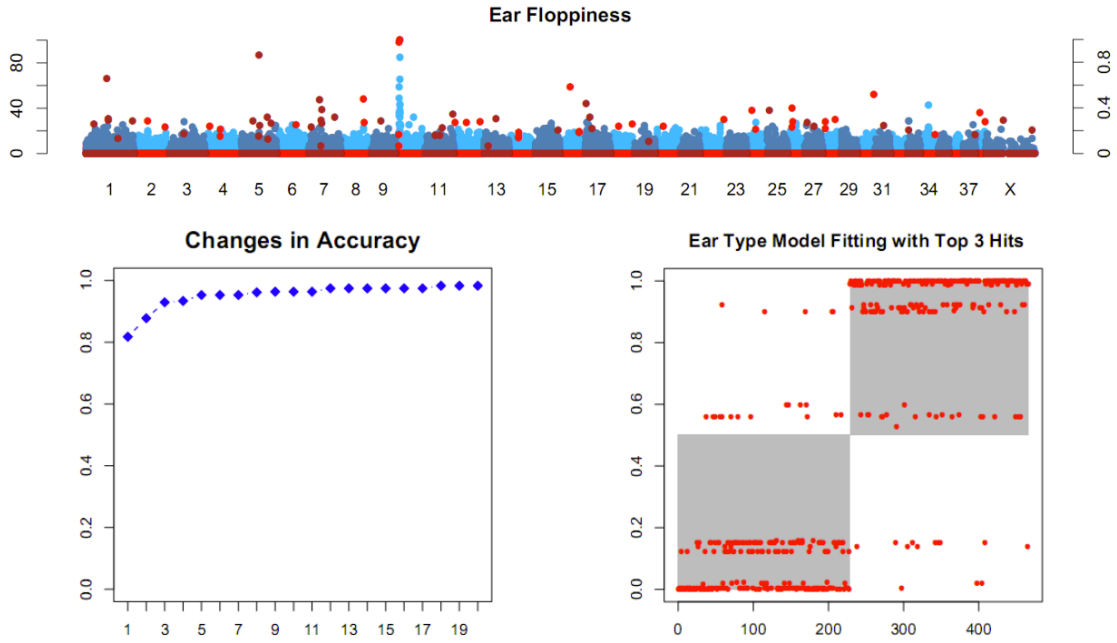

Figure 4.11: Genome-wide scans for ear floppiness and model fitting. The upper panel shows the genome-wide scans, with SMA p-values (in blue) and posteriors from Bayesian scans (in red). The plot on the left in the lower panel shows the changes in accuracy with top hits from the Bayesian scans added into the model sequentially. The plot on the right shows the model fitting with the top 3 hits.

some traits, giving suggested threshold for the posteriors to include terms in the model. For example, a threshold falling in $(0.40, 0.45)$ may be chosen for the trait of height at withers. Such thresholds tend to be different for different traits, so permutations should be done for each individual trait.

Now we consider the permutations based on explained variance. Figure 4.12 shows the results for the traits of body weight, height at withers residual, and outside ear length residual. For each trait, three panels are show in the figure: 1) the adjusted R^2 for models including top 1 to 10 SNPs; 2) the changes in adjusted R^2 when the 1st to 10th SNP is added into the model sequentially; 3) the scatter plot of the changes in R^2 v.s. the R^2 before the term is added. The results shown

Table 4.5: Probabilities in permutations

Trait	t	$P(p > t)$	t	$P(p > t)$
Height at withers	0.45	0.097	0.40	0.107
Height at tail base	0.45	0.094	0.40	0.104
Eye width	0.30	0.089	0.25	0.102
Snout length	0.40	0.096	0.35	0.106
Head length	0.25	0.090	0.20	0.106
Neck length	0.35	0.091	0.30	0.102

in the figure indicate that the top 2 or 3 SNPs for body weight might be claimed to be significant, that the top 1 hit for height at withers residual may be claimed significant, and that the top 2 SNPs seem marginally significant for outside ear length residual.

4.2.6 Breed mapping accounting for breed relatedness

While taking breed relatedness into account, we ran breed mapping on all the traits, and the results were regarded final and reported in [3]. Among all the traits, several traits are especially interesting and are shown in Figure 4.13.

The scan for body weight yields several significant genomic associations, with the six strongest hits occurring at CFA 15.44226659, CFA X.106866624, CFA 10.11440860, CFA X.86813164, CFA 4.42351982, and CFA 7.46842856. The corresponding P-P plot compares the observed distribution of $-\log_{10}(p\text{-value})$ (blue and red dots) to the expected distribution under a model of no-association and demonstrates an excess of significant hits since the tail of the distribution is well

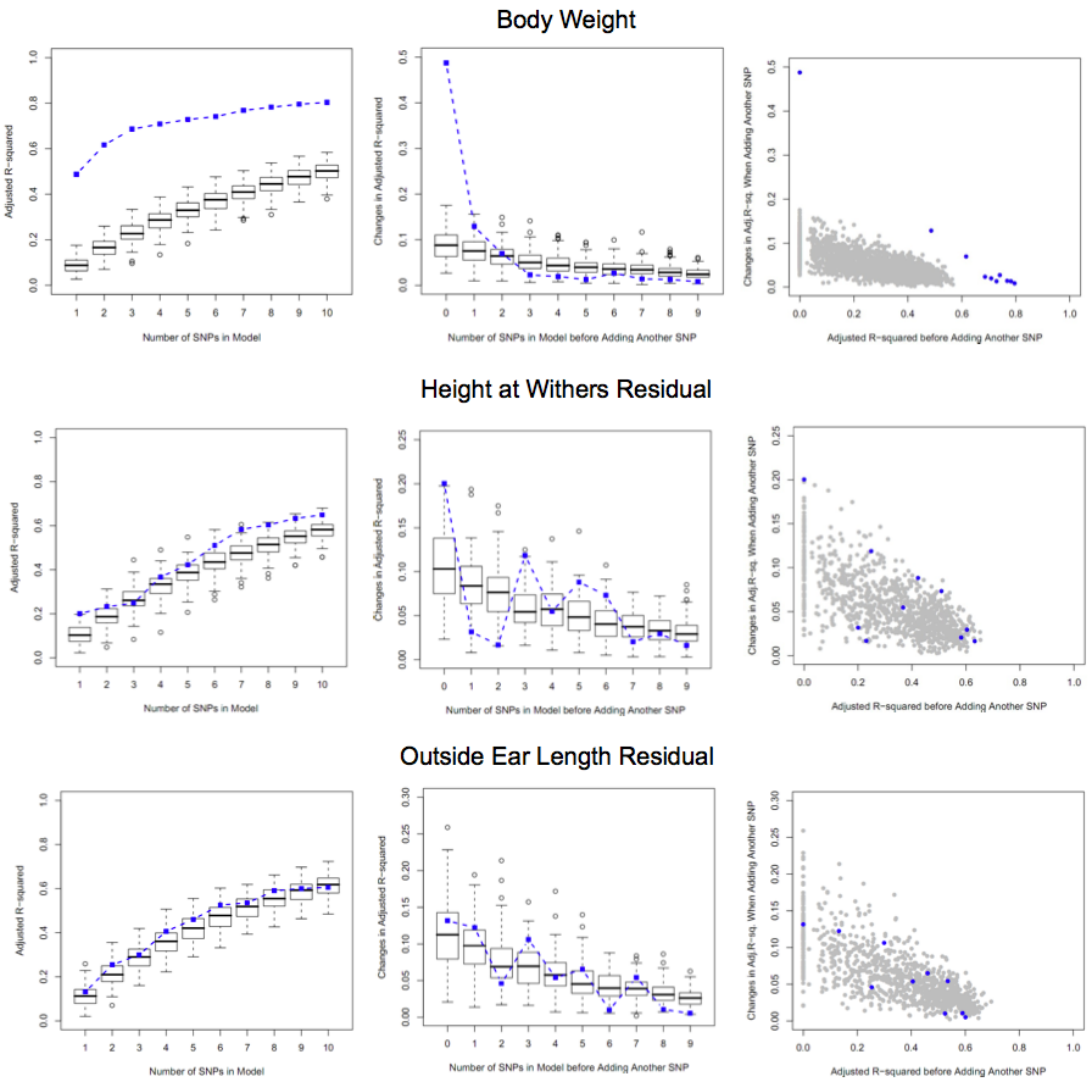

Figure 4.12: Permutation with R^2 for body weight, height at withers residual, and outside ear length residual.

above the diagonal dashed line that represents equality of the expected and observed. When the top six regions (and linked SNPs) are removed, the observed p-value distribution (grey points) is dramatically shifted towards the expected, suggesting these six QTLs clearly account for a portion of the association signal in our data. The first four hits are also among the highest F_{ST} regions in the dog genome [3] with the CFA4 hit also exhibiting elevated $F_{ST} = 0.46$. This is consistent with diversifying selection between breeds for body weight.

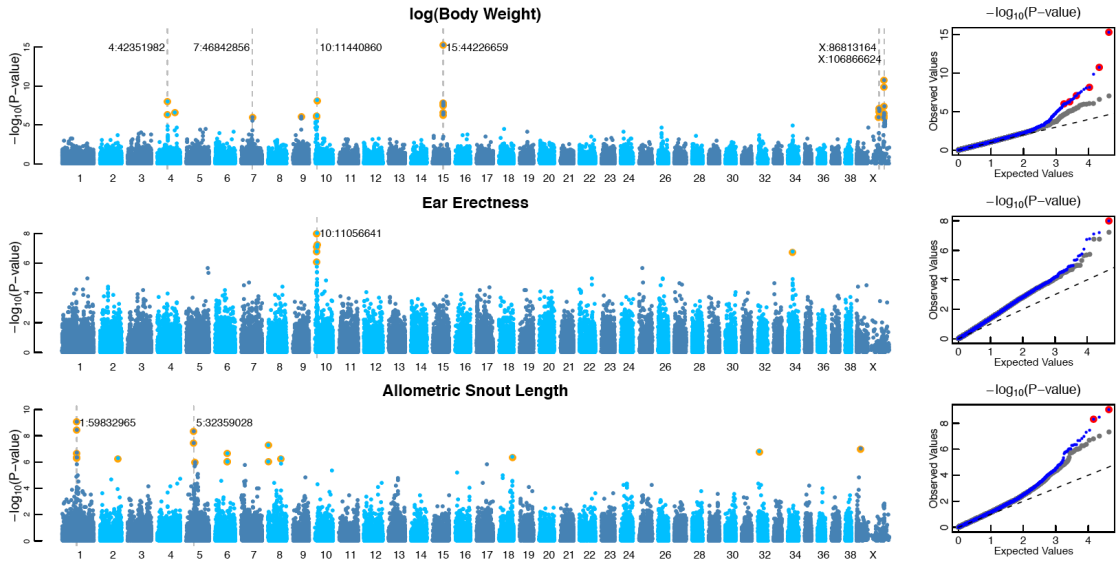

Figure 4.13: Genome-wide association scans across the breeds using allele frequencies of the SNPs and breed-average phenotypes for log(body weight), ear erectness (floppy versus erect ears), and allometric snout length. The P-values of the SNPs were computed using the linear mixed model method for the first and the third traits and using weighted permutation method for ear erectness. SNPs passing Bonferroni's correction are marked with orange circles; SNPs included in best-fit predictive models are marked with gray dashes; the P-P plots for the scans are shown in the right-hand column.

Another key trait is ear type, which also varies substantially among breeds. All adult wild canids have erect ears, but dog breeds are fixed for various ear types, including floppy ears. Treating ear type as a binary variable (floppy v.s. erect), we found in the scan a single region on CFA 10 that is likely responsible for ear floppiness.

Snout length is also a trait that varies considerably between breeds. Breeds scans for absolute snout length return similar genetic regions as those found for body weight, but including $\log\left(\text{body weight}\right)$ as a covariate in the model allows for an allometric correction and reveals QTLs underlying allometric snout

length. The top two hits, CFA 1.59832965 and CFA 5.32359028, are both within the top 5% of high F_{ST} SNPs (F_{ST} = 0.55 and 0.42, respectively) [3].

As a summary of all the associations, Figures 4.14 and 4.15 show the regions with significant hits and their R^2 values (See appendix for the detailed information of these genomic regions). To choose significant hits from all the SNPs, the cutoff for the p-values from breed mapping (accounting for relatedness) is 10^{-5} for absolute tape measurements, and 10^{-4} for absolute bone measurements, and all allometric ones. Then the genomic regions were chosen if they were associated with multiple traits. R^2, computed using the 1-SNP predictive model, is displayed for each trait-region pair with different colors suggesting different magnitudes. When multiple hits are significant in the region, the largest R^2 is used for the R^2 for that region. It can be observed from the figure that correlated traits also have similar patterns of hits.

Using forward stepwise regression, potential hits from the breed mapping are selected to form a multi-SNP predictive model for each trait. Similar to body weight and ear type, most of the measured traits have a model with three or fewer SNPs that can account for most of the variance of the traits. For 55 traits, the mean proportion variance explained by the top 1-, 2-, and 3-SNP models is 0.52, 0.63, and 0.67, respectively, and change to 0.21, 0.32, and 0.4, respectively after controlling for body weight (i.e. considering allometric traits).

Figure 4.16 shows the naive scans that do not control for population structure, and it should be noted that the top genomic regions (Figures 4.14 and 4.15) are similar to this, suggesting the breed relatedness does not seem to bias the association analysis much. Of course, taking into account potential confounding indeed is some protective measure after all.

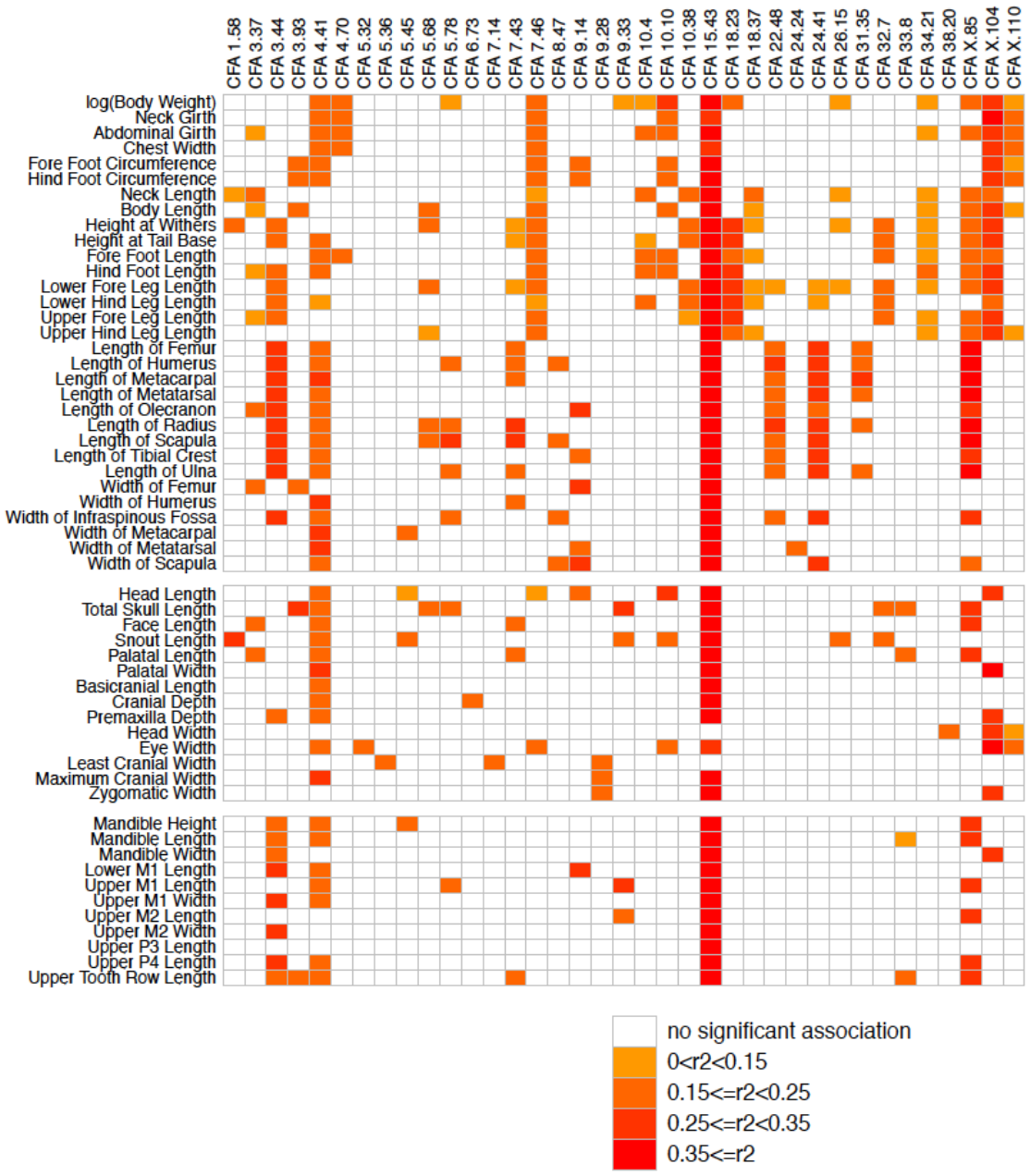

Figure 4.14: Summary of associations between genomic regions and multiple absolute traits. Each row corresponds to an absolute trait, and each column corresponds to a genomic region that has been found associated with multiple traits. The shading of each rectangle shows the magnitude of R^2 statistic of the 1-SNP predictive model for the trait. When multiple SNPs in the region are significant, the largest value of the R^2 statistics is reported.

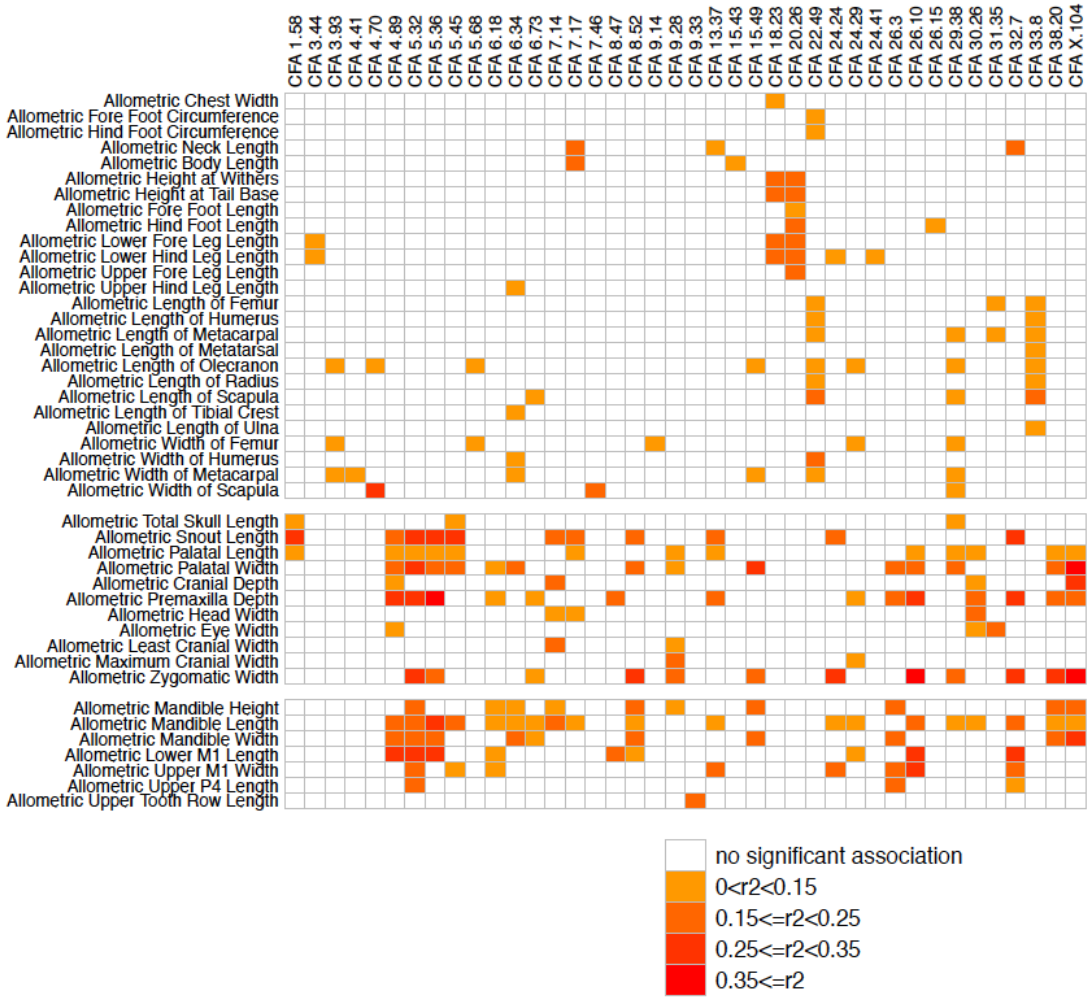

Figure 4.15: Summary of associations between genomic regions and multiple allometric traits. Each row corresponds to an allometric trait, and each column corresponds to a genomic region that has been found associated with multiple traits. The shading of each rectangle shows the magnitude of R^2 statistic of the 1-SNP predictive model for the trait with body weight as a covariate. When multiple SNPs in the region are significant, the largest value of the R^2 statistics is reported.

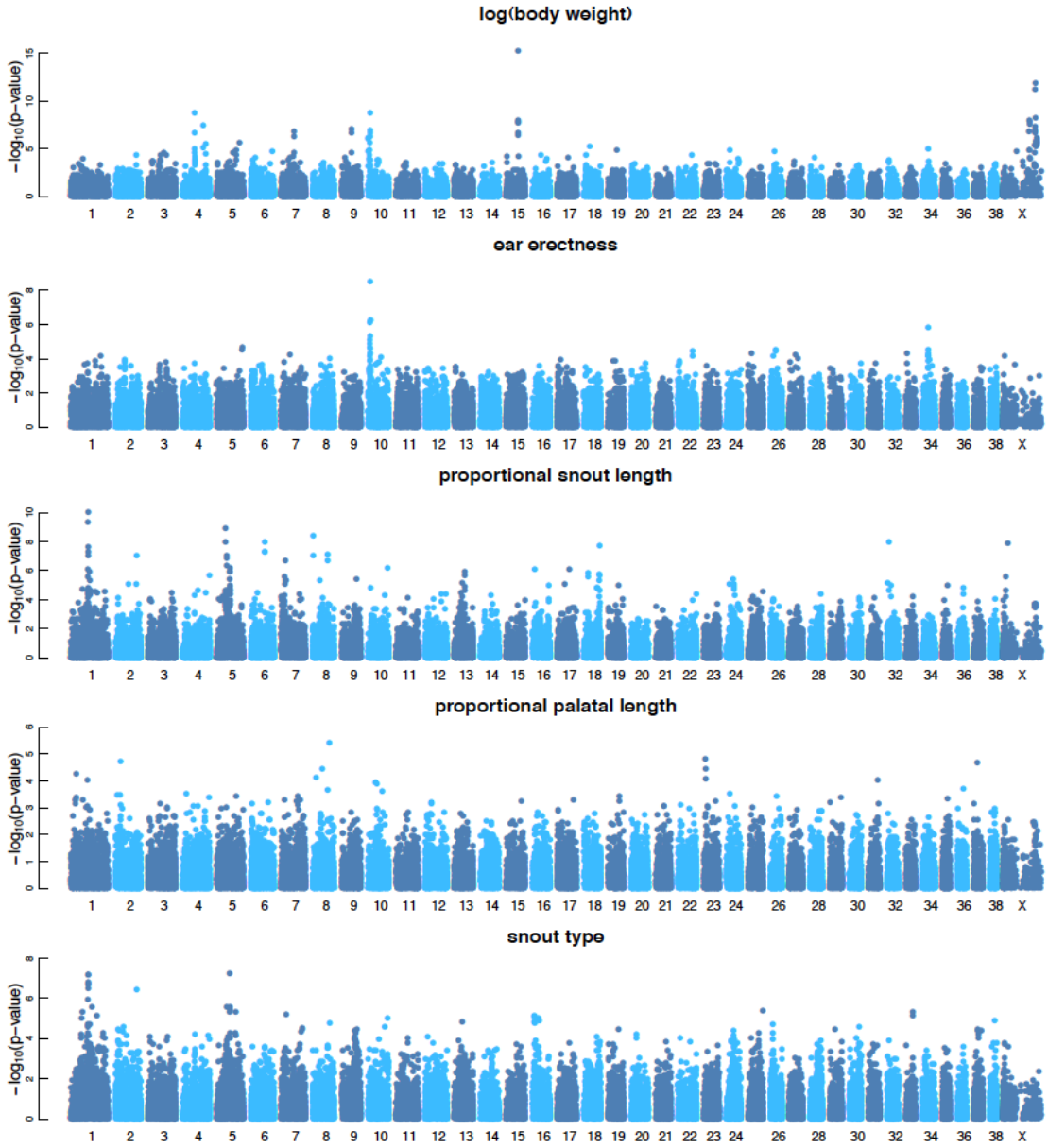

Figure 4.16: Genome-wide association scans using naive tests without accounting for breed relatedness for log(body weight), ear erectness (floppy versus erect ears), proportional snout length, proportional palatal length, and snout type (brachiocephalic versus average)

4.2.7 Model fitting and predictions for body weight

For body weight, we use the top six SNPs associated with breed-average body weight to compute the best-fitting linear predictive models of body weight while ignoring epistasis and non-additive effects at the individual level. We validate the models by making predictions using two samples with known individual weights: 249 dogs from breeds included in CanMap and 50 previously measured outbred village and shelter dogs that were genotyped at the top six SNPs. Table 4.6 shows the coefficients of the SNP effects in the models, where we can find that at least the signs of the coefficients are consistent across all the models.

Table 4.6: Coefficients for the SNP effect(s) in the fitted 1- to 6-SNP models using the top six SNPs. The estimates of the intercept are not shown here. "-" indicates a value is not applicable.

SNP	1-SNP	2-SNP	3-SNP	4-SNP	5-SNP	6-SNP
CFA 15:44226659	0.663	0.551	0.531	0.492	0.445	0.438
CFA X:106866624	–	0.348	0.274	0.207	0.191	0.159
CFA 4:42351982	–	–	0.282	0.270	0.242	0.218
CFA X:86813164	–	–	–	0.211	0.204	0.220
CFA 10:11440860	–	–	–	–	-0.186	-0.171
CFA 7:46842856	–	–	–	–	–	-0.155

Figure 4.17 shows the correlations between predicted phenotypes and observed values. The linear 6-SNP model explains the majority of body weight variation in both the breed dogs and the non-breed village dogs with correlation coefficients of 0.85 and 0.50, respectively. Using the top SNP only (the IGF1 locus), the model explains 50% and 17% of variance in breed and village dogs,

respectively. Using the top 3 SNPs, the model can explain 38% of the variance in village dogs, while the 6-SNP model explains less. This could be a consequence of lower LD in village dogs, or non-genetic factors, the smaller range of body weights in village dogs, or a possibly overfitting of the model. All these results, however, are dramatically different from what have been found in humans.

4.3 Materials and Methods

4.3.1 DNA samples and SNP calling

Blood samples and SNP chips were prepared at the lab of Dr. Elaine Ostrander. Blood samples from dogs were collected with the consent of owners and breeders. DNA was then extracted using a phenol/chloroform standard protocol. Concentration of the DNA was measured using a nanodrop. For the Affymetrix SNP chip, 5ul at a concentration of 50ng/ul was used for each samples. The SNP chip was then run following the Affymetrix GeneChip Mapping 500K 96-well Plate protocol, at the exception of an additional step between the labeling of the samples and the hybridization where we concentrated the samples to be able to load all of it on the SNP chip. [118] In total, 1,659 samples were genotyped on Affymetrix v2 Canine arrays which contain over probes for over 127,000 SNPs markers.

The raw data were then sent to the lab of Dr. Carlos Bustamante to get the genotyping information of each sample. Genotypes were called on 1,400 arrays with highest signal-noise intensity ratios. It was found that the BRLMM-P algorithm yielded approximately 45,000 SNPs that passed quality control fil-

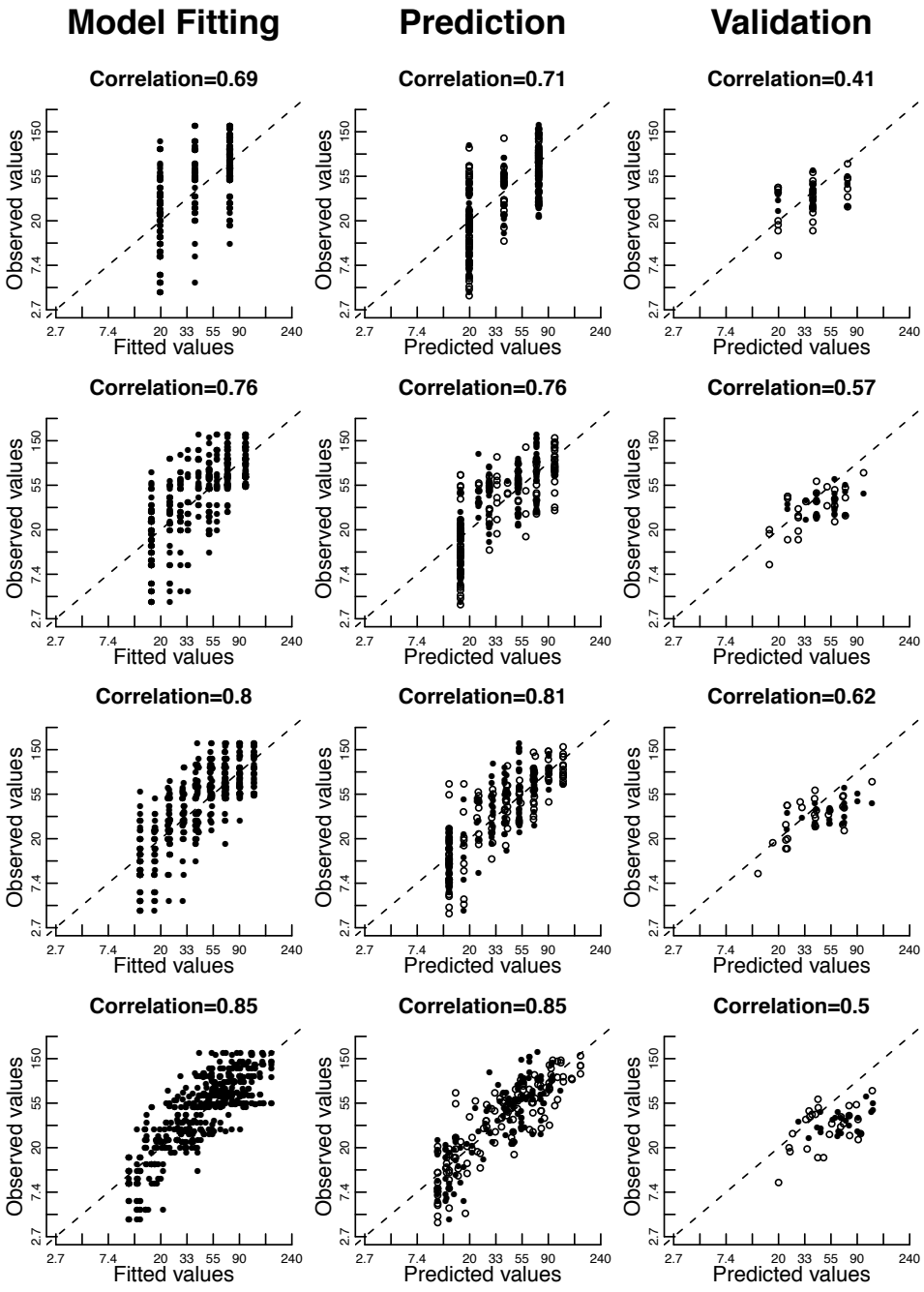

Figure 4.17: Correlation between log(body weight) predicted by the predictive models to the breed-average data (1st column) using breed averages as phenotypes, as well as 249 breed dogs (2nd column) and 50 non-breed village dogs with individual measurements. (A) The predictive model using a single SNP, CFA 15:44226659; (B-D) The predictive models using 2, 3, and 6 top SNPs (in order after CFA 15:44226659, CFA X:106866624, CFA 4:42351982, CFA X:86813164, CFA 10:11440860, and CFA 7.46842856).

tering, and that it consistently over-called heterozygous genotypes. Instead, a novel genotype calling algorithm, MAGIC (Multidimensional Analysis for Genotype Intensity Clustering), developed by Adam Auton and Adam Boyko, was used for SNP calling. On the same 1,400 chips, MAGIC called 60,968 SNPs that passed the quality control filters,. yielding a call rate of 94.6%. After the SNPs were called, the lab of Dr. Ostrander ran several samples in duplicate to verify the accuracy of the call for each SNP. MAGIC gave a concordance rate of over 99.9% for samples run in duplicate. [3]

4.3.2 Sample information and genotype data

The version 2 release of the CanMap data includes genotype information of 1,658 samples at 61,468 SNPs across the 38 autosomes and the X chromosome. There are 1,031 individuals with known breeds and sex (485 males and 546 females, some of which have their sex imputed from genotype data). These individuals are from 80 domestic breeds of dogs, village dogs, as well as jackals, red wolves, wolves, and coyotes, each of which has around 10-12 individuals.

All the genetic analyses were based on this data set, started with necessary preprocessing and filtering. Depending on the analyses, the filtering can be removing duplicates of samples, removing samples with certain missingness of genotyping, and removing SNPs with low call rates, low minor allele frequencies, and/or high missingness of genotyping.

4.3.3 Traits under investigation and phenotypic values

One important trait under investigation is body weight, with which a SNP has been found strongly associated [2]. Now with the densely genotyped SNP data, we revisited the trait in this project to see if we could replicate previous results and identify additional associations. Among those genotyped in CanMap, 249 dogs have individual body weights. There are also 50 village dogs with genotypes and individual body weights available [117]. For other dogs with known breeds, we used average body weights of male dogs in AKC breeds [108] which were treated as breed characteristics and assigned to each individual of the same breed as phenotypic values. Sex adjustment was estimated using known sex and individual phenotypes to adjust the body weights of female dogs.

When blood samples were collected from dogs, owners and breeders also were asked through questionnaires for tape measurements of different external parts of dogs. These external measurements (e.g. height at withers, body length, etc., see Table 4.7 and Figure 4.18) were considered as morphological traits, each with more than 1,000 observations. However, each of these traits has only around 200 dogs with individual measurements in the CanMap data set, and most observations do not have corresponding genotype information. Similar to body weight, we computed the breed average for each trait using the dogs older than one year old if at least two observations were available for the trait. The breed averages were treated as breed characteristics and hence used as phenotypic values of individual dogs.

Bone measurements (skull and limb measurements) were taken from the museum specimens of dogs [106, 107]. Table 4.8 lists the names of the traits under investigation and Figures B.3 and B.4 show how the measurements were

Table 4.7: Tape measurements asked for in the questionnaires and taken by dog owners and breeders. The numbers correspond to different parts of the dog shown in Figure 4.18 which were actually measured externally. Several measurements, e.g. hind foot length, fore foot length, etc., were taken on both sides, the averages of which were used for the phenotypic values.

No.	Trait Name	No.	Trait Name
1	Height at withers	12	Abdominal girth
2	Height at base of tail	13	Chest width
3	Head width	14	Hind foot length
4	Eye width	15	Hind foot circumference
5	Snout length	16	Lower hind leg length
6	Head length	17	Upper hind leg length
7	Neck length	18	Fore foot length
8	Body length	19	Fore foot circumference
9	Tail length	20	Lower fore leg length
10	Ear length	21	Upper fore leg length
11	Neck girth		

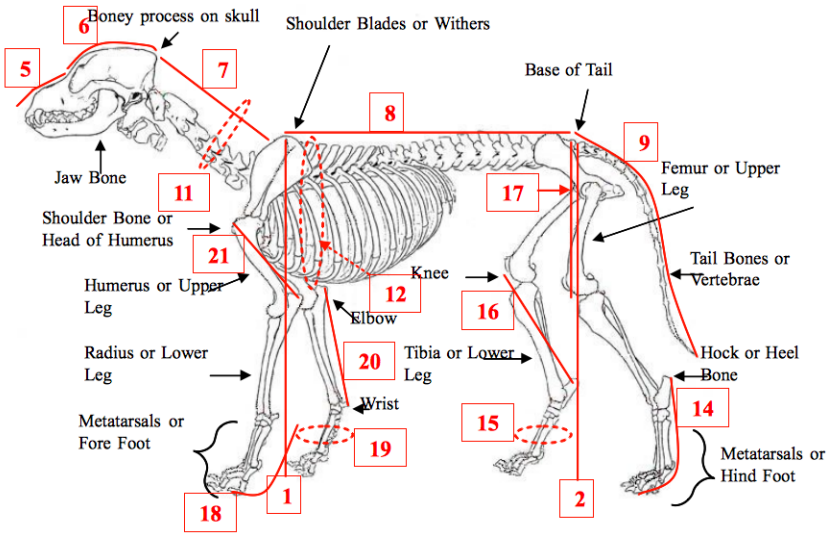

Figure 4.18: The skeleton of a generic dog. The diagram was in the questionnaires for dog owners and breeders, helping them take the tape measurements. Red lines suggest the tape measurements that have been taken (Table 4.7). Note that the measurements were taken externally, although the diagram shows the skeleton. This diagram was reproduced from [3].

taken on skulls and skeletons (reproduced from [106, 107] with permissions from the author). Since no genotype information is available for these samples, the breed averages of bone measurements were computed and used as phenotypic values.

Table 4.8: Trait names of skull and limb measurements, taken from museum specimens of dogs.

Abbreviation	Trait Name	Abbreviation	Trait Name
TSL	Total skull length	MH	Mandible height
FL	Face length	ML	Mandible length
PL	Palatal length	M_1L	Lower M1 length
BSL	Basisphenoid length	BCL	Basicranial length
BOL	Basioccipital length	LF	Length of femur
PW	Palatal width	LH	Length of humerus
TRL	Upper tooth row length	WH	Width of humerus
P^3L	Upper P3 length	WF	Width of femur
P^4L	Upper P4 length	LR	Length of radius
M^1L	Upper M1 length	LMC	Length of metacarpal
M^2L	Upper M2 length	WMC	Width of metacarpal
M^1W	Upper M1 width	LTC	Length of tibial crest
M^2W	Upper M2 width	LS	Length of scapula
MCW	Maximum cranial width	WS	Width of scapula
ZW	Zygomatic width	WSF	Width of infraspinous fossa
LCW	Least cranial width	LU	Length of ulna
CD	Cranial depth	LO	Length of olecranon
PD	Premaxilla depth	LMT	Length of metatarsal
MW	Mandible width	WMT	Width of metatarsal

At certain stage of the project, we also considered some ratio traits that were

derived from the external measurements in Table 4.7. Table 4.9 lists each of the traits and how the phenotypic value for a breed was calculated from breed averages of external measurements.

Table 4.9: Ratios of external measurements. The ratios, following similar definition in previous studies [1], can be used to compare relative body sizes of different breeds.

Ratio	Calculation	Ratio	Calculation
Head ratio	Head length/Body length	Chest ratio	Chest width/Body length
Snout ratio	Snout length/Head length	Eye ratio	Eye width/Head width
Leg ratio	Leg length/Body length	Ear ratio	Ear length/Head length
Tail ratio	Tail length/Body length	Girth ratio	Neck girth/Chest girth
Neck ratio	Neck length/Body length	Foot ratio	Foot circumference/Chest girth

4.3.4 Population structure and breed relatedness

In our GWA studies on the morphological traits, sample relatedness and breed relatedness may be potential confounding factors. For example, if a set of small dog breeds are more closely related than large dog breeds, then it is possible that the loci identified simply are associated with historical relatedness rather than body size. To address this problem, we carried out principal component analysis (PCA) [38] and also studied similarity matrices of identity by state (IBS). PCA, carried out by coauthors in [3], evaluated population structure using 5,157 unlinked SNPs genotyped on 890 dogs from 80 breeds. It is observed that breed groups do not tend to form clusters, and single breeds or pairs of closely related breeds are "pulled out" as one adds PC dimensions. To quantify sample relatedness in a similarity matrix, an individual-by-individual IBS matrix was

calculated using PLINK [78]. Breed relatedness is quantified by taking the averages of elements in the individual IBS matrix within breeds that leads to a breed-average IBS matrix. Figure 4.19 shows the breed-average IBS matrix for 79 breeds that were studied for body weight.

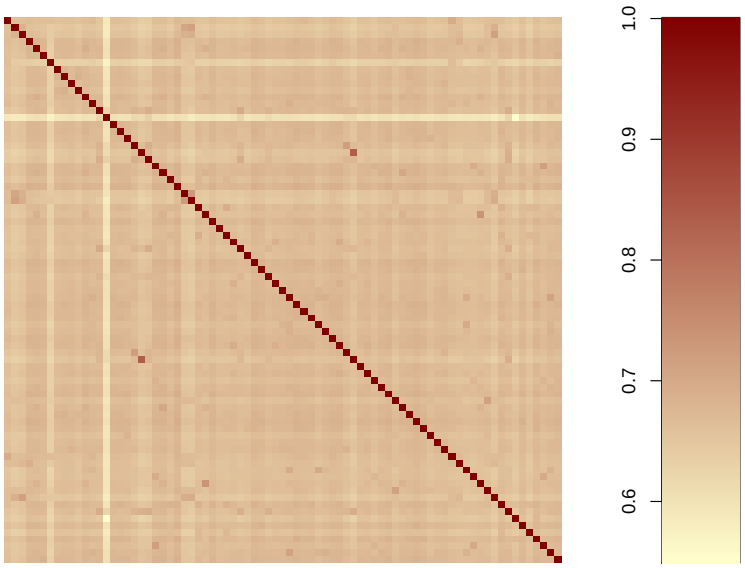

Figure 4.19: The breed-average IBS matrix including 79 breeds studied for body weight. Each element of the matrix is the average IBS between individuals from two breeds, calculated from the individual-by-individual IBS matrix.

4.3.5 Individual mapping and breed mapping

To search for associations with a trait, two kinds of analyses were considered, termed individual mapping and breed mapping. Individual mapping takes breed averages as the phenotypic values of individuals and relates them to the genotypes. This is similar to what is usually done in GWA studies, except that the with-in breed phenotypic variation is overlooked as the information is not

available. Such analyses were considered in our initial studies on the traits. With most individuals in the CanMap data set having breed averages as their phenotypic values, it makes more sense to do breed mapping, i.e. relating the breed averages to allele frequencies. If the breed mapping is considered, then the sample size is the number of breeds involved, e.g. 79 for body weight. It should be noted that the summary of allele frequencies overlooks the uncertainty in genotypes. The final results of our GWA studies are based on breed mapping, and were reported here in this chapter and in [3]. Breed mapping was also used in previous studies [1, 119].

4.3.6 Single marker analysis (Naive scans)

Single marker analysis tests one SNP at a time for association, and the term "naive scan" is used since no population structure is taken into account. If the genotypes of each individual are used, PLINK is used for testing associations with both binary traits and quantitative traits. P-values are calculated based on asymptotic distributions of the test statistics. If the allele frequencies of each individual are used, either linear regression or logistic regression with only one SNP involved at a time is considered.

4.3.7 Bayesian regression using mixture priors (Bayesian scans)

Bayesian linear models [63] and generalized linear models (Chapter 2) with mixture priors are considered in the initial analysis of the GWA studies. Either genotypes or allele frequencies are used as explanatory variables, and Bayesian

regressions are applied to both individual mapping and breed mapping. However, we are not considering a random effect to model genetic relatedness, although it is straightforward to add an additional term to the model.

4.3.8 Linear mixed model (LMM scans)

For continuous traits, a linear mixed model [1] was used to test each of the SNPs for association while also controlling for breed relatedness. A random effect was assumed to follow a multivariate normal distribution with mean 0 and the correlation matrix being the breed-average IBS matrix. The fixed effect takes one SNP at a time, and a covariate can be added for allometric traits, for example, log(breed average body weight) for external measurement traits, or log(breed average total skull length) for skeletal traits.

4.3.9 Weighted bootstrap method (WB scans)

For binary traits, a weighted bootstrap method was used to test each of the SNPs for association. The method was developed by Keyan Zhao, a coauthor of [3]. The phenotypes were bootstrapped with weights accounting for breed relatedness, and the empirical distributions of test statistics were obtained for calculating p-values. Suppose the sample size is N and denote the IBS matrix as K with the value between breed i and breed j equals to K_{ij}. A typical bootstrap step works as follows: with i iterating from 1 to N, a phenotype for the ith breed was assigned with that of the jth breed, j being chosen with a probability of $K_{ij}/\sum_j K_{ij}$. Then a χ^2 correlation test-statistic is obtained at each bootstrap

step between the bootstrapped phenotypes and the allele frequencies for each SNP. The p-values can be obtained by calculating the frequencies that the test-statistics in bootstrap are larger than that from observed phenotypes.

I was responsible for applying this method to the analyses of the binary traits.

4.3.10 Threshold for claiming significance

In linear regression, when more terms are added in the model, it tends to overfit the data. For example, in Figure 4.10, with more terms added in the model, the BIC keeps dropping and the correlation keeps increasing. The linear Bayesian mixture model, lacking of metrics similar to p-values, needs a way to evaluate the significance of its hits, as without a validation set, it is difficult to decide how many terms should be included in the model.

In order to draw a line for including the terms and avoiding overfitting, we tried some permutation methods. One permutation method is to try to obtain a threshold for the posterior probabilities. This is how we did the permutation. First, calculate the average phenotypes μ_{breed} and standard deviation σ_{breed} for each breed; second, for each permutation:

1. Randomly switch breed labels. Those dogs of the same breed *breed* are still in the same breed, but are assigned with a different breed label $breed^*$;
2. Assign each dog with draws from $N\left(\mu_{breed^*}, \sigma^2_{breed^*}\right)$, where the mean and standard deviation are the values of the assigned label $breed^*$;

3. Rerun the Bayesian scan with the new phenotypes and the top 100 SNPs in the original run.

Do 1,000 runs of such permutation, and then calculate the probability that the posterior exceeds the threshold $P(p > t)$, where p is the posterior of a SNP having a non-negligible effect, and t is a certain probability threshold.

The second strategy of permutation is to look at the variance explained by multiple SNPs. For each permutation,

1. Randomly switch breed labels.
2. Assign breed averages to individual phenotypes
3. Run Bayesian scans to find the top n hits
4. Fit a series of linear model using the top 1 to n hits and retain the adjusted R^2
5. Repeat 1-4 for N times (e.g. N = 100, 200 or larger)

4.3.11 Modeling fitting and validation

Genotypes of individual dogs and phenotypes taking breed averages were used for building the predictive models based on genotypes. The SNPs prioritized by the scans accounting for breed relatedness were chosen and further selected for best predictive models using forward stepwise regressions.

For those traits with individual measurements, we used the predictive models to predict the individual phenotypes and compared them to the observed

values. The predictive models for body weights were also validated on a data set of 50 village dogs with individual body weights [117].

4.4 Discussion

The CanMap project generates a high density map of canine genetic variation by genotyping $\sim 1,000$ dogs from 80 domestic dog breeds, 83 wild canids and 10 outbred African shelter dogs across $\sim 60,000$ SNPs. Using this genomic resource, we are able to investigate the genetic basis underlying the tremendous morphological diversity in the domestic dog. We carried out GWA analyses on complex traits including body weight, external body dimensions, cranial, dental and long bone shape and size with and without allometric scaling. For most of the traits, a small number of QTLs (≤ 3) are found to explain the majority of phenotypic variation (often $> 70\%$). At least for body weight, they also seem to account for phenotypic variation within outbred village dogs. The association scan results of the dominance of a few genes with large effects offer a sharp contrast to recent findings in humans, where GWA studies on quantitative traits such as height, weight, BMI, blood pressure suggest the most phenotypic variation in our species is governed by a large number of mutations with small effects. This helps enrich our understanding of the process of domestication and artificial selection in the domestic dog. Boyko et al. [3] also find that regions associated with morphological variation account for at least the eleven top F_{ST} regions across dog breeds, consistent with both strong selection for morphology and a simplified genetic architecture for these quantitative traits in dogs. The novel genomic approaches using the domestic dog as a model system, as in the CanMap project, enrich our understanding of what impact strong directional se-

lection has had on the genetic architecture of complex traits known to be under selection.

The applications of statistical methods for associations are described in detail in this chapter. The morphological traits in dogs have larger variation across breeds than within breeds, and the phenotypic values for most individuals are not available. Given the special data structure, various methods have been tried, and finally emphasis was put on the methods that map breed averages of phenotypes with allele frequencies of genotypes. Both principal component analysis and mixed-effect models have been considered to account for population structure, mostly possible genetic relatedness between breeds. Although the final results are presented using single marker analysis, i.e. testing one SNP at a time, several multi-locus methods have also been considered, largely for a purpose of exploration. Some of these methods are covered in previous chapters, while other are not. The results seem similar to those obtained from single marker analysis, suggesting at least these methods have similar performance. There are also some novelty in applying these multi-locus methods, for example, the attempts to evaluate the significance of the obtained signals. Currently all the traits are studied one at a time, and it seems to be the proper thing to do given the modest sample size. At the aspect of methodology, however, it would be interesting to consider a multivariate response for each individual containing several correlated traits. Follow-up investigations on these statistical problems are well worthwhile and will be considered in future research.

APPENDIX A

DETAILS OF THE ALGORITHMS FOR BAYESIAN MIXTURE MODELS

A.1 Logistic mixture model with adaptive independence sampler

Let $\boldsymbol{\beta}_{-j}$ be $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_M)'$ with β_j excluded, $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{iM})'$, $\boldsymbol{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_N)'$, and $\boldsymbol{X}_{i,-j}$ be $\boldsymbol{X}_i$ with X_{ij} excluded, $\boldsymbol{Y} = (Y_1, \cdots, Y_N)'$. $[A|B]$ denotes the conditional distribution of A given B, and $[A]$ denotes the marginal distribution of A. For initialization of the Gibbs sampler, μ and $\boldsymbol{\beta}$ can be initialized by the estimates from the logistic regression on one SNP at a time, or by small random numbers. Missing genotype data are initialized by allele frequencies given the value of Y. A typical iteration step of the MCMC algorithm includes the following subroutines.

a. Impute missing genotype data: Sample each missing genotype value X_{ij} from its full conditional posterior distribution

$$[X_{ij}|Y_i, \boldsymbol{X}, \mu, \boldsymbol{\beta}] \propto [Y_i|X_{ij}, \boldsymbol{X}, \mu, \beta] \times [X_{ij}|\boldsymbol{X}_{i,-j}]$$

b. Update μ: A new μ is sampled from its full conditional posterior distribution. Sampling μ is achieved by the Metropolis-Hastings algorithm as follows:

1. Define $P(\mu|\beta, Y) = \frac{e^{-\mu\sum_i(1-y_i)}}{\prod_i(1+e^{-\mu-\sum_j\beta_j x_{ij}})}$ which is in scale of the full conditional distribution of μ.
2. Draw $\mu^* \sim N(\mu, \sigma_\mu^2)$ where σ_μ^2 is initialized to be 1.

3. Compute α using

$$\alpha(\mu, \mu^*) = \min\left(1, \frac{P(\mu^*|\boldsymbol{\beta}, Y)}{P(\mu|\boldsymbol{\beta}, Y)} \cdot \frac{N(\mu|\mu^*, \sigma_\mu^2)}{N(\mu^*|\mu, \sigma_\mu^2)}\right) = \min\left(1, \frac{P(\mu^*|\boldsymbol{\beta}, Y)}{P(\mu|\boldsymbol{\beta}, Y)}\right)$$

4. Draw $u \sim U(0,1)$, and if $u < \alpha(\mu, \mu^*)$, update μ by μ^*, otherwise no update

σ_μ^2 is tuned every certain number of iteration steps: If the acceptance rate $\gamma \leq 0.4$, $\sigma_{\mu,new}^2 = \sigma_\mu^2(2 - \frac{\gamma}{0.4})^{-2}$; if $\gamma \geq 0.6$, $\sigma_{\mu,new}^2 = \sigma_\mu^2(2 - \frac{1-\gamma}{0.4})^2$.

c. Update $\boldsymbol{\beta}$: Each β_j of $\boldsymbol{\beta}$ is sampled from its full conditional posterior distribution. Sampling β_j is achieved by the proposed independence adaptive sampler as follows.

1. Define $P(\beta_j|\mu, \boldsymbol{\beta}_{-j}, Y, \mu_+, \sigma_+^2, \mu_-, \sigma_-^2)$ as the term at the right hand side of equation (2.11), which is in scale of the full conditional posterior distribution of β_j.
2. Sample β_j^* from the proposal distribution $q(\beta_j, \beta_j^*)$, defined by equation (2.12), and the transition probability function $q(s,t), (s,t \in \{+,-,0\})$ is defined in Figure 2.4.
3. Compute $\alpha(\beta_j, \beta_j^*) = \min\left(1, \frac{P(\beta_j^*|\mu, \boldsymbol{\beta}_{-j}, Y, \mu_+, \sigma_+^2, \mu_-, \sigma_-^2)}{P(\beta_j|\mu, \boldsymbol{\beta}_{-j}, Y, \mu_+, \sigma_+^2, \mu_-, \sigma_-^2)} \cdot \frac{q(\beta_j^* \to \beta_j)}{q(\beta_j \to \beta_j^*)}\right)$
4. Draw $u \sim U(0,1)$, and if $u < \alpha(\beta_j, \beta_j^*)$, update β_j by β_j^*, otherwise no update.

The above procedures go through $j = 1, \cdots, m$.

d. Update p_+ and p_-: Sample p_+, p_- from the full conditional distribution

$$[p_{\beta+}, p_{\beta-}, 1 - p_{\beta+} - p_{\beta-})|\beta] \sim \text{Dirichlet}(m_+^{(0)} + \tilde{n}_+, m_-^{(0)} + \tilde{n}_-, m_0^{(0)} + M - \tilde{n}_+ - \tilde{n}_-)$$

where $\left(m_+^{(0)}, m_-^{(0)}, m_0^{(0)}\right)$ can be set to be $(1,1,1)$ and $\tilde{n}_+$, $\tilde{n}_-$ are the numbers of positive β_j's and negative β_j's, respectively. p_+ and p_- are both truncated

to be in the range of $(0, \min(m_0/M, 1))$, and m_0 is a user-defined parameter in the order of $\sqrt{N}$.

e. Update $\mu_+, \sigma_+^2, \mu_-, \sigma_-^2$: The pair (μ_+, σ_+^2) is updated using positive β_j's, while the pair (μ_-, σ_+^2) is updated using negative β_j's, both of which use conjugate prior distributions (2.9) and (2.10). The details of this part can be found in [66]

σ_+^2 is sampled from its marginal posterior distribution, given positive β's:

$$[\sigma_+^2|\boldsymbol{\beta}] \sim \text{Inv-}\chi^2(\nu_{+n}, \sigma_{+n}^2) \tag{A.1}$$

μ_+ is sampled from its conditional posterior distribution, given positive β's:

$$[\mu_+|\sigma_+^2, \beta] \sim N(\mu_{+n}, \sigma_+^2/\kappa_{+n}) \tag{A.2}$$

where

$$\begin{aligned}
\mu_{+n} &= \frac{\kappa_{+0}}{\kappa_{+0}+\tilde{n}_+}\mu_{+0} + \frac{\tilde{n}_+}{\kappa_{+0}+\tilde{n}_+}\bar{\beta}_+ \\
\kappa_{+n} &= \kappa_{+0} + \tilde{n}_+ \\
\nu_{+n} &= \nu_{+0} + \tilde{n}_+ \\
\sigma_{+n}^2 &= \frac{\nu_{+0}}{\nu_{+n}}\sigma_{+0}^2 + \frac{\tilde{n}_+ - 1}{\nu_{+n}}s_+^2 + \frac{\kappa_{+0}\tilde{n}_+}{\nu_{+n}(\kappa_{+0}+\tilde{n}_+)}(\bar{\beta}_+ - \mu_{+0})^2
\end{aligned}$$

and $\bar{\beta}_+$ is the mean of positive β_j's, s_+^2 is the sample variance estimate with the factor $\frac{1}{\tilde{n}_+ - 1}$. In our implementation, the hyper-parameters are set as follows: if $\tilde{n}_+ = 0$, $\mu_{+0} = 0$, $\kappa_{+0} = 1$, $\nu_{+0} = 3$, and $\sigma_{+0}^2 = 0.0625$; otherwise, μ_{+0}, κ_{+0} and σ_{+0}^2 are set to be the sample mean, total number, and variance of initial β_j's that are positive, $\nu_{+0} = 3$.

The sampling of (μ_-, σ_-^2) follows the similar procedure by replacing the subscript "+" with "−" in (A.1) and (A.2).

Steps a)-e) are repeated until the chain reaches its stationary.

A.2 Probit mixture model with data augmentation

Besides the notations in the previous section, let $\boldsymbol{Z} = (Z_1, \cdots, Z_n)'$. For initialization of the Gibbs sampler, μ and $\boldsymbol{\beta}$ can be initialized by the estimates from the probit regression on one SNP at a time, or by small random numbers. Missing genotype data are initialized by allele frequencies given the value of Y. A typical iteration step of the Gibbs sampler includes the following subroutines.

a. Sample latent variable $Z_i, (i = 1, \cdots, n)$: Sample Z_i from its full conditional posterior distribution

$$[Z_i|\boldsymbol{X}, \boldsymbol{Y}, \mu, \boldsymbol{\beta}] \propto N_+(\mu + \sum_i X_{ij}\beta_j, 1)I_{\{Y_i=1\}} + N_-(\mu + \sum_i X_{ij}\beta_j, 1)I_{\{Y_i=0\}}$$

b. Impute missing genotype data: Sample each missing genotype value X_{ij} from its full conditional posterior distribution

$$[X_{ij}|Z_i, \boldsymbol{X}_{i,-j}, \mu, \boldsymbol{\beta}] \propto [Z_i|X_{ij}, \boldsymbol{X}_{i,-j}, \mu, \boldsymbol{\beta}] \times [X_{ij}|\boldsymbol{X}_{i,-j}]$$

c. Update μ: Sample μ from its full conditional posterior distribution

$$\mu|\boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{\beta} \sim N\left(\frac{1}{N}\sum_i (Z_i - \sum_j X_{ij}\beta_j), \frac{1}{N}\right)$$

d. Update $\beta_j, (j = 1, \cdots, m)$: Sample each β_j from its full conditional distribution. For $j = 1, \cdots, m$,

$$\beta_j|\boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{\beta}_{-j}, p_+, p_-, \sigma_+^2, \sigma_-^2 \sim \tilde{p}_{j+}N_+(\tilde{\mu}_{j+}, \tilde{\sigma}_{j+}^2) + \tilde{p}_{j-}N_-(\tilde{\mu}_{j-}, \tilde{\sigma}_{j-}^2) + \tilde{p}_{j0}I_{\{\beta_j=0\}}$$

where

$$\tilde{\mu}_{j+} = \frac{\sigma_+^2 \sum_{i=1}^n x_{ij}(z_i - \mu - \sum_{l\neq j}\beta_l x_{il})}{1 + \sigma_+^2 \sum_{i=1}^n x_{ij}^2}, \quad \tilde{\sigma}_{j+}^2 = \frac{\sigma_+^2}{1 + \sigma_+^2 \sum_{i=1}^n x_{ij}^2}$$

$$\tilde{\mu}_{j-} = \frac{\sigma_-^2 \sum_{i=1}^n x_{ij}(z_i - \mu - \sum_{l \neq j} \beta_l x_{il})}{1 + \sigma_-^2 \sum_{i=1}^n x_{ij}^2}, \quad \tilde{\sigma}_{j-}^2 = \frac{\sigma_-^2}{1 + \sigma_-^2 \sum_{i=1}^n x_{ij}^2}$$

$$\tilde{p}_{j+} = \frac{2p_+ \frac{\tilde{\sigma}_{j+}}{\sigma_+} \Phi\left(\frac{\tilde{\mu}_{j+}}{\tilde{\sigma}_{j+}}\right) \exp\left(\frac{\tilde{\mu}_{j+}^2}{2\tilde{\sigma}_{j+}^2}\right)}{1 - p_+ - p_- + 2p_+ \frac{\tilde{\sigma}_{j+}}{\sigma_+} \Phi\left(\frac{\tilde{\mu}_{j+}}{\tilde{\sigma}_{j+}}\right) \exp\left(\frac{\tilde{\mu}_{j+}^2}{2\tilde{\sigma}_{j+}^2}\right) + 2p_- \frac{\tilde{\sigma}_{j-}}{\sigma_-} \Phi\left(\frac{\tilde{\mu}_{j-}}{\tilde{\sigma}_{j-}}\right) \exp\left(\frac{\tilde{\mu}_{j-}^2}{2\tilde{\sigma}_{j-}^2}\right)}$$

$$\tilde{p}_{j-} = \frac{2p_- \frac{\tilde{\sigma}_{j-}}{\sigma_-} \Phi\left(\frac{\tilde{\mu}_{j-}}{\tilde{\sigma}_{j-}}\right) \exp\left(\frac{\tilde{\mu}_{j-}^2}{2\tilde{\sigma}_{j-}^2}\right)}{1 - p_+ - p_- + 2p_+ \frac{\tilde{\sigma}_{j+}}{\sigma_+} \Phi\left(\frac{\tilde{\mu}_{j+}}{\tilde{\sigma}_{j+}}\right) \exp\left(\frac{\tilde{\mu}_{j+}^2}{2\tilde{\sigma}_{j+}^2}\right) + 2p_- \frac{\tilde{\sigma}_{j-}}{\sigma_-} \Phi\left(\frac{\tilde{\mu}_{j-}}{\tilde{\sigma}_{j-}}\right) \exp\left(\frac{\tilde{\mu}_{j-}^2}{2\tilde{\sigma}_{j-}^2}\right)}$$

$\tilde{p}_{j0} = 1 - \tilde{p}_{j+} - \tilde{p}_{j-}$, and p_+,p_-, σ_+^2 and σ_-^2 are the same in the prior (2.5) at the current step.

e. Sample p_+ and p_-: Sample p_+, p_- from the full conditional distribution

$$[p_{\beta+}, p_{\beta-}, 1 - p_{\beta+} - p_{\beta-})|\boldsymbol{\beta}] \sim \text{Dirichlet}(m_+^{(0)} + \tilde{n}_+, m_-^{(0)} + \tilde{n}_-, m_0^{(0)} + M - \tilde{n}_+ - \tilde{n}_-)$$

where $\left(m_+^{(0)}, m_-^{(0)}, m_0^{(0)}\right)$ is usually set to be $(1, 1, 1)$ and $\tilde{n}_+$, $\tilde{n}_-$ are the numbers of positive β_j's and negative β_j's, respectively. p_+ and p_- are both truncated to be in the range of $(0, \min(m_0/M, 1))$, and m_0 is a user-defined parameter in the order of $\sqrt{N}$.

f. Sample σ_+^2 and σ_-^2: Sample σ_+^2 and σ_-^2 from the full conditional distributions

$$\sigma_+^{-2}|\boldsymbol{\beta} \sim \Gamma\left(1 + \frac{\tilde{n}_+}{2}, (1 + \sum_j \beta_j^2 I_{\{\beta_j > 0\}})^{-1}\right)$$

$$\sigma_-^{-2}|\boldsymbol{\beta} \sim \Gamma\left(1 + \frac{\tilde{n}_-}{2}, (1 + \sum_j \beta_j^2 I_{\{\beta_j < 0\}})^{-1}\right)$$

Steps a)-f) are repeated until the chain reaches its stationary.

A.3 Logistic mixture model with data augmentation

Let $\boldsymbol{Z} = (Z_1, \cdots, Z_n)'$. For initialization of the Gibbs sampler, μ and $\boldsymbol{\beta}$ can be initialized by the estimates from the logistic regression on one SNP at a time,

or by small random numbers. Missing genotype data are initialized by allele frequencies given the value of Y. A typical iteration step of the Gibbs sampler includes the following subroutines.

a. Sample latent variable $Z_i, (i = 1, \cdots, N)$: Sample Z_i from its full conditional posterior distribution

$$[Z_i|\boldsymbol{X}, \boldsymbol{Y}, \mu, \boldsymbol{\beta}, \boldsymbol{\phi}] \propto N_+(\mu + \sum_i X_{ij}\beta_j, \phi^{-1})I_{\{Y_i=1\}} + N_-(\mu + \sum_i X_{ij}\beta_j, \phi^{-1})I_{\{Y_i=0\}}$$

b. Impute missing genotype data: Sample each missing genotype value X_{ij} from its full conditional posterior distribution

$$[X_{ij}|Z_i, \boldsymbol{X}_{i,-j}, \mu, \boldsymbol{\beta}] \propto [Z_i|X_{ij}, \boldsymbol{X}_{i,-j}, \mu, \boldsymbol{\beta}] \times [X_{ij}|\boldsymbol{X}_{i,-j}]$$

c. Update μ: Sample μ from its full conditional posterior distribution

$$\mu|\boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{\beta} \sim N\left(\frac{1}{N}\sum_i (Z_i - \sum_j X_{ij}\beta_j), \frac{1}{N}\phi^{-1}\right)$$

d. Update $\beta_j, (j = 1, \cdots, M)$: Sample each β_j from its full conditional distribution. For $j = 1, \cdots, M$,

$$\beta_j|\boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{\beta}_{-j}, p_+, p_-, \sigma_+^2, \sigma_-^2 \sim \tilde{p}_{j+}N_+(\tilde{\mu}_{j+}, \tilde{\sigma}_{j+}^2) + \tilde{p}_{j-}N_-(\tilde{\mu}_{j-}, \tilde{\sigma}_{j-}^2) + \tilde{p}_{j0}I_{\{\beta_j=0\}}$$

where

$$\tilde{\mu}_{j+} = \frac{\sigma_+^2 \sum_{i=1}^n x_{ij}(z_i - \mu - \sum_{l\neq j}\beta_l x_{il})}{1 + \sigma_+^2 \sum_{i=1}^n x_{ij}^2}, \quad \tilde{\sigma}_{j+}^2 = \frac{\sigma_+^2}{1 + \sigma_+^2 \sum_{i=1}^n x_{ij}^2}$$

$$\tilde{\mu}_{j-} = \frac{\sigma_-^2 \sum_{i=1}^n x_{ij}(z_i - \mu - \sum_{l\neq j}\beta_l x_{il})}{1 + \sigma_-^2 \sum_{i=1}^n x_{ij}^2}, \quad \tilde{\sigma}_{j-}^2 = \frac{\sigma_-^2}{1 + \sigma_-^2 \sum_{i=1}^n x_{ij}^2}$$

$$\tilde{p}_{j+} = \frac{2p_+\frac{\tilde{\sigma}_{j+}}{\sigma_+}\Phi\left(\frac{\tilde{\mu}_{j+}}{\tilde{\sigma}_{j+}}\right)\exp\left(\frac{\tilde{\mu}_{j+}^2}{2\tilde{\sigma}_{j+}^2}\right)}{1 - p_+ - p_- + 2p_+\frac{\tilde{\sigma}_{j+}}{\sigma_+}\Phi\left(\frac{\tilde{\mu}_{j+}}{\tilde{\sigma}_{j+}}\right)\exp\left(\frac{\tilde{\mu}_{j+}^2}{2\tilde{\sigma}_{j+}^2}\right) + 2p_-\frac{\tilde{\sigma}_{j-}}{\sigma_-}\Phi\left(\frac{\tilde{\mu}_{j-}}{\tilde{\sigma}_{j-}}\right)\exp\left(\frac{\tilde{\mu}_{j-}^2}{2\tilde{\sigma}_{j-}^2}\right)}$$

$$\tilde{p}_{j-} = \frac{2p_- \frac{\tilde{\sigma}_{j-}}{\sigma_-} \Phi\left(\frac{\tilde{\mu}_{j-}}{\tilde{\sigma}_{j-}}\right) \exp\left(\frac{\tilde{\mu}_{j-}^2}{2\tilde{\sigma}_{j-}^2}\right)}{1 - p_+ - p_- + 2p_+ \frac{\tilde{\sigma}_{j+}}{\sigma_+} \Phi\left(\frac{\tilde{\mu}_{j+}}{\tilde{\sigma}_{j+}}\right) \exp\left(\frac{\tilde{\mu}_{j+}^2}{2\tilde{\sigma}_{j+}^2}\right) + 2p_- \frac{\tilde{\sigma}_{j-}}{\sigma_-} \Phi\left(\frac{\tilde{\mu}_{j-}}{\tilde{\sigma}_{j-}}\right) \exp\left(\frac{\tilde{\mu}_{j-}^2}{2\tilde{\sigma}_{j-}^2}\right)}$$

$\tilde{p}_{j0} = 1 - \tilde{p}_{j+} - \tilde{p}_{j-}$, and p_+,p_-, σ_+^2 and σ_-^2 are the same in the prior (2.5) at the current step.

e. Sample p_+ and p_-: Sample p_+, p_- from the full conditional distribution

$$[p_{\beta+}, p_{\beta-}, 1 - p_{\beta+} - p_{\beta-})|\beta] \sim \text{Dirichlet}(m_+^{(0)} + \tilde{n}_+, m_-^{(0)} + \tilde{n}_-, m_0^{(0)} + M - \tilde{n}_+ - \tilde{n}_-)$$

where $\left(m_+^{(0)}, m_-^{(0)}, m_0^{(0)}\right)$ is usually set to be $(1, 1, 1)$ and $\tilde{n}_+$, $\tilde{n}_-$ are the numbers of positive β_j's and negative β_j's, respectively. p_+ and p_- are both truncated to be in the range of $(0, \min(m_0/M, 1))$, and m_0 is a user-defined parameter in the order of $\sqrt{N}$.

f. Sample σ_+^2 and σ_-^2: Sample σ_+^2 and σ_-^2 from the full conditional distributions

$$\sigma_+^{-2}|\boldsymbol{\beta} \sim \Gamma\left(1 + \frac{\tilde{n}_+}{2}, (1 + \sum_j \beta_j^2 I_{\{\beta_j > 0\}})^{-1}\right)$$

$$\sigma_-^{-2}|\boldsymbol{\beta} \sim \Gamma\left(1 + \frac{\tilde{n}_-}{2}, (1 + \sum_j \beta_j^2 I_{\{\beta_j < 0\}})^{-1}\right)$$

Steps a)-f) are repeated until the chain reaches its stationary.

APPENDIX B

SUPPLEMENTARY MATERIALS FOR THE DOG STUDY

B.1 Decay of linkage disequilibrium

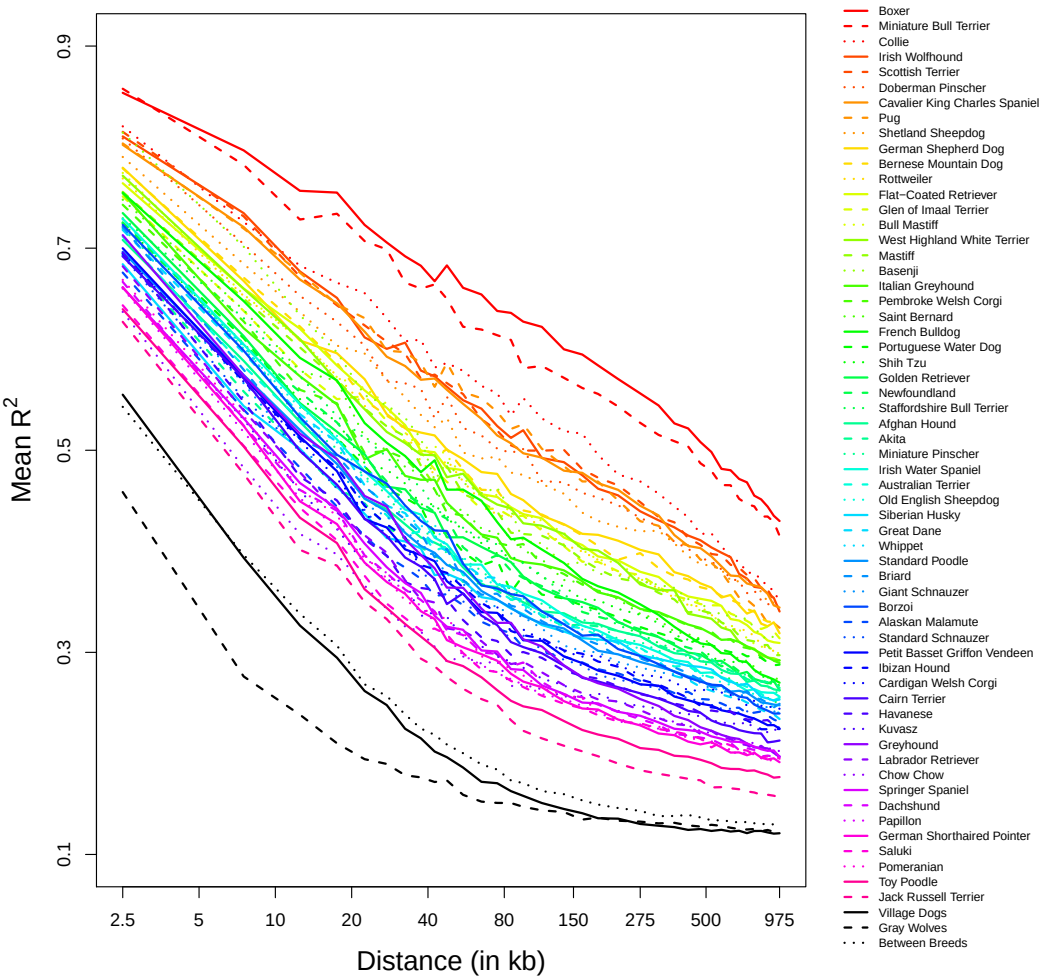

Figure B.1: Analysis of 10 individuals from each of 59 breeds and one population of village dogs and wolves: LD decay curves based on mean R^2, including mean LD decay when dogs are selected from different breeds ("Between breeds"). Calculated by coauthors in [3] and plotted by me.

B.2 Distributions of long runs of homozygosity

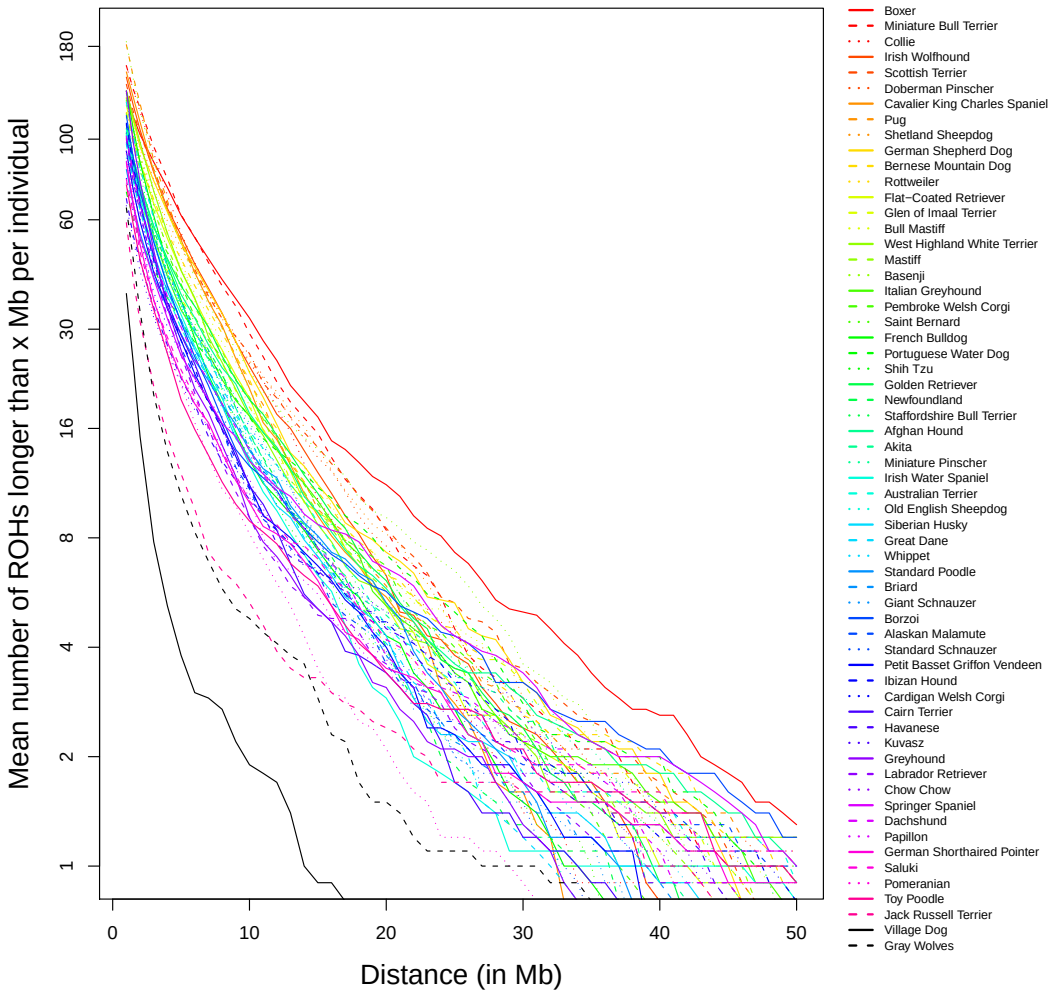

Figure B.2: Analysis of 10 individuals from each of 59 breeds and one population of village dogs and wolves: distribution of long runs of homozygosity in each group. Calculated by coauthors in [3] and plotted by me.

B.3 Illustration of traits

Figures B.3 and B.4 are reproduced from [106, 107] permitted by the author.

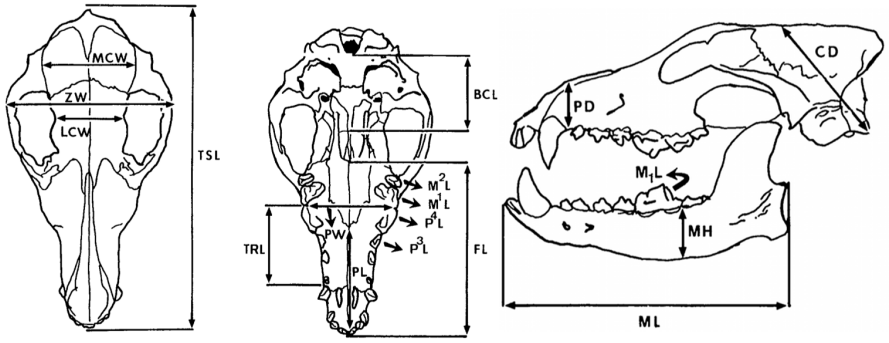

Figure B.3: Skull measurements taken on the museum specimens. The diagram shows how the measurements were taken, and the trait names are shown in Table 4.8.

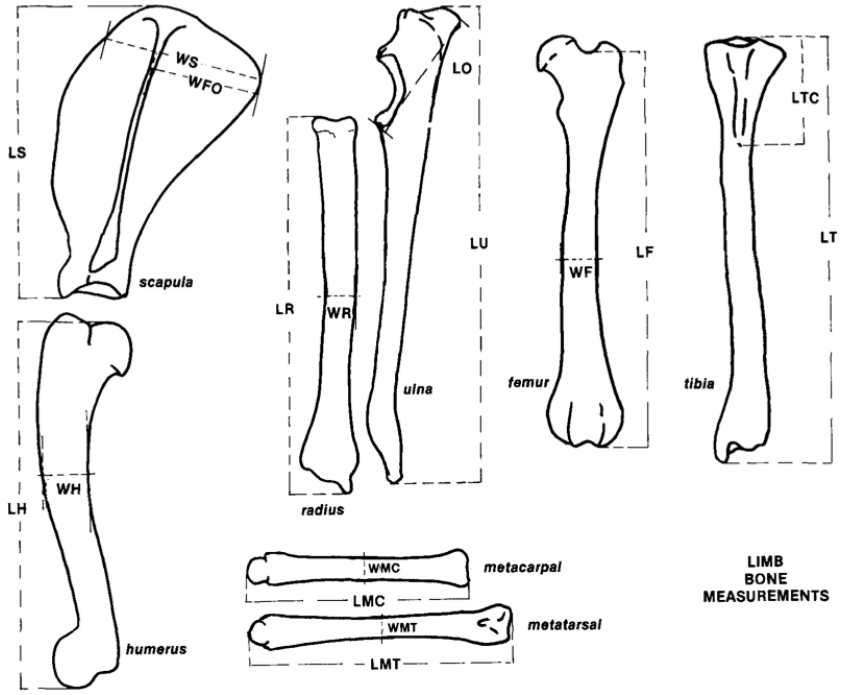

Figure B.4: Limb measurements taken on the museum specimens. The diagram shows how the measurements were taken, and the trait names are shown in Table 4.8.

B.4 Genomic regions associated with multiple morphological traits

Table B.1: The start and end position of the genomic regions shown Figures 4.14 and 4.15. The first and the fourth columns show the labels of the regions, including the chromosome information.

Region	Start	End	Region	Start	End
CFA 1.58	58,901,153	60,560,116	CFA 10.38	38,887,016	38,893,195
CFA 3.37	37,862,000	37,915,954	CFA 13.37	37,887,773	38,311,889
CFA 3.44	44,099,822	44,615,782	CFA 15.43	43,829,663	44,273,435
CFA 3.93	93,135,743	93,851,186	CFA 15.49	49,357,128	49,482,238
CFA 4.41	41,649,432	42,610,286	CFA 18.23	23,298,242	23,298,242
CFA 4.70	70,170,855	70,224,679	CFA 18.37	37,597,915	37,600,713
CFA 4.89	89,694,270	91,046,967	CFA 20.26	26,188,392	26,188,392
CFA 5.32	32,186,026	34,783,508	CFA 22.48	48,487,967	48,487,967
CFA 5.36	36,166,554	37,709,895	CFA 22.49	49,880,005	52,091,127
CFA 5.45	45,047,863	47,828,053	CFA 24.24	24,786,870	24,921,352
CFA 5.68	68,156,340	68,630,173	CFA 24.29	29,855,461	31,190,526
CFA 5.78	78,904,012	78,904,012	CFA 24.41	41,706,206	41,706,206
CFA 6.18	18,878,983	18,878,983	CFA 26.3	3,685,160	3,958,609
CFA 6.34	34,874,368	36,590,724	CFA 26.10	10,197,126	12,323,801
CFA 6.73	73,648,849	74,665,975	CFA 26.15	15,658,589	16,300,140
CFA 7.14	14,348,713	15,771,494	CFA 29.38	38,599,844	39,950,394
CFA 7.17	17,473,677	17,816,157	CFA 30.26	26,478,076	27,853,518
CFA 7.43	43,583,214	44,059,025	CFA 31.35	35,859,072	35,926,711
CFA 7.46	46,837,936	47,080,158	CFA 32.7	7,158,539	8,795,512
CFA 8.47	47,903,947	48,796,876	CFA 33.8	8,724,130	10,390,145
CFA 8.52	52,981,891	53,000,452	CFA 34.21	21,437,482	21,437,482
CFA 9.14	14,967,697	15,745,898	CFA 38.20	20,239,198	20,239,198
CFA 9.28	28,700,337	28,700,337	CFA X.0	673,261	673,261
CFA 9.33	33,918,647	35,320,954	CFA X.85	85,867,623	88,183,292
CFA 10.4	4,913,212	5,157,275	CFA X.104	104,724,717	108,201,633
CFA 10.10	10,859,628	11,440,860	CFA X.110	110,332,086	110,850,946

BIBLIOGRAPHY

[1] Jones, P., Chase, K., Martin, A., Davern, P., Ostrander, E. A., and Lark, K. G. (2008). Single-nucleotide-polymorphism-based association mapping of dog stereotypes. Genetics *179*, 1033–1044.

[2] Sutter, N., Bustamante, C., Chase, K., Gray, M., Zhao, K., Zhu, L., Padhukasahasram, B., Karlins, E., Davis, S., and Jones, P. (2007). A single igf1 allele is a major determinant of small size in dogs. Science *316*, 112.

[3] Boyko, A. R., Quignon, P., Li, L., Schoenebeck, J., Degenhardt, J., Lohmueller, K., Brisbin, A., Parker, H. G., VonHoldt, B. M., Cargill, M., et al. (2010). A simple genetic architecture underlies quantitative traits in dogs. Manuscript in revision.

[4] Venter, J., Adams, M., Myers, E., Li, P., and Mural, R. (2001). The sequence of the human genome. Science *291*, 1304.

[5] Lander, E., Linton, L., Birren, B., Nusbaum, C., and Zody, M. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860.

[6] Collins, F. S., Morgan, M., and Patrinos, A. (2003). The human genome project: lessons from large-scale biology. Science *300*, 286–90.

[7] Gibbs, R., Belmont, J., Hardenbol, P., Willis, T., and Yu, F. (2003). The international hapmap project. Nature *426*, 789.

[8] International HapMap Consortium. (2005). A haplotype map of the human genome. Nature *437*, 1299–320.

[9] Leppert, M., Dixon, M., Peiffer, A., Qiu, R., Kent, A., and Kato, K. (2007). A second generation human haplotype map of over 3.1 million snps. Nature *449*, 851–862.

[10] Kaiser, J. (2008). Dna sequencing. a plan to capture human diversity in 1000 genomes. Science *319*, 395.

[11] Yu, J., Hu, S., Wang, J., Wong, G., Li, S., Liu, B., and Deng, Y. (2002). A draft sequence of the rice genome (oryza sativa l. ssp. indica). Science *296*, 79–92.

[12] Dean, R., Yu, Y., Zharkikh, A., Shen, R., and Sahasrabudhe, S. (2002). A draft sequence of the rice genome (oryza sativa l. ssp. japonica). Science pp. 92–100.

[13] Hillier, L., Miller, W., Birney, E., and Warren, W. (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature *432*, 695–716.

[14] Lindblad-Toh, K., Wade, C., Mikkelsen, T., and Karlsson, E. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature *438*, 803–819.

[15] Chimpanzee Sequencing and Analysis Consortium. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. Nature *437*, 69–87.

[16] Rhesus Macaque Genome Sequencing and Analysis Consortium, Gibbs, R. A., Rogers, J., Katze, M. G., Bumgarner, R., Weinstock, G. M., Mardis, E. R., Remington, K. A., Strausberg, R. L., Venter, J. C., et al. (2007). Evolutionary and biomedical insights from the rhesus macaque genome. Science *316*, 222–34.

[17] The Bovine Genome Sequencing and Analysis Consortium, Elsik, C. G., Tellam, R. L., and Worley, K. C. (2009). The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science *324*, 522–528.

[18] Iyengar, S. and Elston, R. (2007). The genetic basis of complex traits: rare variants or "common gene, common disease"? Methods in molecular biology *376*, 71.

[19] Lander, E. S. and Schork, N. J. (1994). Genetic dissection of complex traits. Science *265*, 2037–48.

[20] Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. Science *273*, 1516–1517.

[21] Lander, E. (1996). The new genomics: global views of biology. Science *274*, 536–539.

[22] Christensen, K. and Murray, J. (2007). What genome-wide association studies can do for medicine. New England Journal of Medicine *356*, 1094–1097.

[23] Edwards, A. O., Ritter, R., Abel, K. J., Manning, A., Panhuysen, C., and Farrer, L. A. (2005). Complement factor h polymorphism and age-related macular degeneration. Science *308*, 421–4.

[24] Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., et al. (2005). Complement factor h polymorphism in age-related macular degeneration. Science *308*, 385–9.

[25] Haines, J. L., Hauser, M. A., Schmidt, S., Scott, W. K., Olson, L. M., Gallins, P., Spencer, K. L., Kwan, S. Y., Noureddine, M., Gilbert, J. R., et al. (2005). Complement factor h variant increases the risk of age-related macular degeneration. Science *308*, 419–21.

[26] The Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661–678.

[27] Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A. M., Hu, T. T., et al. (2010). Genome-wide association study of 107 phenotypes in arabidopsis thaliana inbred lines. Nature pp. doi:10.1038/nature08800.

[28] Dickson, S., Wang, K., Krantz, I., and Hakonarson, H. (2010). Rare variants create synthetic genome-wide associations. ncbi.nlm.nih.gov *8*, e1000294.

[29] Hirschhorn, J. and Daly, M. (2005). Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics *6*, 95–108.

[30] McCarthy, M., Abecasis, G., Cardon, L., Goldstein, D., Little, J., Ioannidis, J., and Hirschhorn, J. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Reviews Genetics *9*, 356–369.

[31] Pearson, T. and Manolio, T. (2008). How to interpret a genome-wide association study. Journal of the American Medical Association *299*, 1335–1344.

[32] Altshuler, D., Daly, M. J., and Lander, E. S. (2008). Genetic mapping in human disease. Science *322*, 881–8.

[33] Newman, D. L., Abney, M., McPeek, M. S., Ober, C., and Cox, N. J. (2001). The importance of genealogy in determining genetic associations with complex traits. American journal of human genetics *69*, 1146–8.

[34] Helgason, A., Yngvadóttir, B., Hrafnkelsson, B., Gulcher, J., and Stefánsson, K. (2005). An icelandic example of the impact of population structure on association studies. Nature genetics *37*, 90–5.

[35] Abiola, O., Angel, J. M., Avner, P., Bachmanov, A. A., Belknap, J. K., Bennett, B., Blankenhorn, E. P., Blizard, D. A., Bolivar, V., Brockmann, G. A., et al. (2003). The nature and identification of quantitative trait loci: a community's view. Nature Reviews Genetics *4*, 911–6.

[36] Devlin, B. and Roeder, K. (1999). Genomic control for association studies. Biometrics *55*, 997–1004.

[37] Pritchard, J., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics *155*, 945–959.

[38] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet *38*, 904–9.

[39] Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet *38*, 203–8.

[40] Zhao, K., Aranzana, M., Kim, S., and Lister, C. (2007). An Arabidopsis example of association mapping in structured samples. PLoS Genetics *3*, 71–82.

[41] Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. Genetics *178*, 1709–23.

[42] Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-Y., Freimer, N. B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nat Genet *42*, 348–54.

[43] Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A.,

Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. Nature Genetics *42*, 355–360.

[44] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc. B. *58*, 267–288.

[45] Candes, E. and Tao, T. (2007). The dantzig selector: statistical estimation when p is much larger than n. Annals of Statistics *35*, 2313–2351.

[46] Xu, S. (2003). Estimating polygenic effects using markers of the entire genome. Genetics *163*, 789–801.

[47] George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. Journal of the American Statistical Association *88*, 881–889.

[48] Hoggart, C. J., Whittaker, J. C., Iorio, M. D., and Balding, D. J. (2008). Simultaneous analysis of all snps in genome-wide and re-sequencing association studies. PLoS Genetics *4*, e1000130.

[49] Huang, H., Eversley, C. D., Threadgill, D. W., and Zou, F. (2007). Bayesian multiple quantitative trait loci mapping for complex traits using markers of the entire genome. Genetics *176*, 2529–2540.

[50] Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics *25*, 714–21.

[51] Logsdon, B., Hoffman, G., and Mezey, J. (2010). A variational bayes algorithm for fast and accurate multiple locus genome-wide association analysis. BMC bioinformatics *11*.

[52] Parker, H., Kukekova, A., Akey, D., Goldstein, O., Kirkness, E., Baysac, K., Mosher, D., Aguirre, G., Acland, G., and Ostrander, E. (2007). Breed relationships facilitate fine-mapping studies: A 7.8-kb deletion cosegregates with collie eye anomaly across multiple dog breeds. Genome Research *17*, 1562.

[53] Shearin, A. L. and Ostrander, E. A. (2010). Canine morphology: hunting for genes and tracking mutations. PLoS Biol *8*, e1000310.

[54] Akaike, H. (1973). Information theory and an extension of the maximum

likelihood principle. In Petrov, B. N. and Csaki, F., eds., Second International Symposium on Information Theory pp. 267–281.

[55] Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics *6*, 461–464.

[56] Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association *101*, 1418–1429.

[57] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association *96*, 1348–1360.

[58] James, G., Radchenko, P., and Lv, J. (2009). DASSO: connections between the Dantzig selector and lasso. J. Roy. Statist. Soc. Ser. B *71*, 127–142.

[59] Mitchell, T. and Beauchamp, J. (1988). Bayesian variable selection in linear regression. Journal of the American Statistical Association *83*, 1023–1032.

[60] Genkin, A., Lewis, D. D., and Madigan, D. (2006). Large-scale bayesian logistic regression for text categorization. Technical report.

[61] Yi, N., George, V., and Allison, D. B. (2003). Stochastic search variable selection for identifying multiple quantitative trait loci. Genetics *164*, 1129–1138.

[62] Yi, N. and Xu, S. (2008). Bayesian LASSO for quantitative trait loci mapping. Genetics *179*, 1045–1055.

[63] Zhang, M., Montooth, K., Wells, M., Clark, A., and Zhang, D. (2005). Mapping multiple quantitative trait loci by bayesian classification. Genetics *169*, 2305–2318.

[64] Zhang, M., Zhang, D., and Wells, M. (2008). Variable selection for large p small n regression models with incomplete data: mapping QTL with epistases. BMC bioinformatics *9*.

[65] Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. Journal of the Royal Statistical Society. Series A (General) *135*, 370–384.

[66] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). Bayesian Data Analysis (Ed. 2). (Chapman & Hall/CRC).

[67] Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions and the bayesian restoration of images. IEEE Trans. Pattn. Anal. Mach. Intel. *6*, 721–741.

[68] Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. (Morgan Kaufmann Publishers, Inc.).

[69] Gao, H., Williamson, S., and Bustamante, C. D. (2007). A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. Genetics *176*, 1635–1651.

[70] Wright, S. (1934). An analysis of variablility in number of digits in an inbred strain of guinea pigs. Genetics *19*, 506–536.

[71] Kinney, S. and Dunson, D. (2007). Fixed and random effects selection in linear and logistic models. Biometrics *63*, 690–698.

[72] Taylor, J. and Verbyla, A. (2004). Joint modelling of location and scale parameters of the t distribution. Statistical Modelling *4*, 91–112.

[73] Scheet, P. and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. The American Journal of Human Genetics *78*, 629–644.

[74] Hudson, R. R. (2002). Generating samples under a wright-fisher neutral model of genetic variation. Bioinformatics *18(2)*, 337–338.

[75] Nelson, M. R., Bryc, K., King, K. S., Indap, A., Boyko, A. R., Novembre, J., Briley, L. P., Maruyama, Y., Waterworth, D. M., Waeber, G., et al. (2008). The population reference sample, popres: a resource for population, disease, and pharmacological genetics research. The American Journal of Human Genetics *83*, 347–358.

[76] Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. The Proceedings of the 23rd International Conference on Machine Learning pp. 233–240.

[77] Zhang, Y. and Liu, J. S. (2007). Bayesian inference of epistatic interactions in case-control studies. Nat Genet *39*, 1167–1173. 10.1038/ng2110.

[78] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M., Bender,

D., Maller, J., Sklar, P., Bakker, P. D., Daly, M., et al. (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics *81*, 559–575.

[79] Su, W.-L., Sieberts, S. K., Kleinhanz, R. R., Lux, K., Millstein, J., Molony, C., and Schadt, E. E. (2010). Assessing the prospects of genome-wide association studies performed in inbred mice. Mamm Genome *21*, 143–52.

[80] Snelling, W. M., Allan, M. F., Keele, J. W., Kuehn, L. A., McDaneld, T., Smith, T. P. L., Sonstegard, T. S., Thallman, R. M., and Bennett, G. L. (2010). Genome-wide association study of growth in crossbred beef cattle. Journal of animal science *88*, 837–48.

[81] Voight, B. F. and Pritchard, J. K. (2005). Confounding from cryptic relatedness in case-control association studies. PLoS Genet *1*, e32.

[82] Hey, J. and Machado, C. (2003). The study of structured populations—new hope for a difficult and divided science. Nature Reviews Genetics *4*, 535–543.

[83] Aranzana, M. J., Kim, S., Zhao, K., Bakker, E., Horton, M., Jakob, K., Lister, C., Molitor, J., Shindo, C., Tang, C., et al. (2005). Genome-wide association mapping in arabidopsis identifies previously known flowering time and pathogen resistance genes. PLoS Genetics *1*, e60.

[84] Buckler, E. S., Holland, J. B., Bradbury, P. J., Acharya, C. B., Brown, P. J., Browne, C., Ersoz, E., Flint-Garcia, S., Garcia, A., Glaubitz, J. C., et al. (2009). The genetic architecture of maize flowering time. Science *325*, 714–8.

[85] Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., et al. (2008). Genes mirror geography within europe. Nature *456*, 98–101.

[86] Zhao, K., Nordborg, M., and Marjoram, P. (2007). Genome-wide association mapping using mixed-models: application to GAW15 Problem 3. BMC proceedings *1*, S164.

[87] Orelien, J. G. and Edwards, L. J. (2008). Fixed-effect variable selection in linear mixed models using R^2 statistics. Computational Statistics & Data Analysis *52*, 1896 – 1907.

[88] Lavergne, C., Martinez, M.-J., and Trottier, C. (2008). Empirical model selection in generalized linear mixed effects models. Computational Statistics *23*, 99–109.

[89] Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. Journal of the American Statistical Association *99*, 710–723.

[90] Ni, X., Zhang, D., and Zhang, H. H. (2009). Variable selection for semiparametric mixed models in longitudinal studies. Biometrics *1*, 79–88.

[91] Aulchenko, Y., de Koning, D., and Haley, C. (2007). Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics *177*, 577–585.

[92] Amin, N., Duijn, C. V., and Aulchenko, Y. (2007). A genomic background based method for association analysis in related individuals. PLoS One *2*.

[93] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. Annals of Statistics *32*, 407–451.

[94] Henderson, C. R. (1950). Estimation of genetic parameters. The Annals of Mathematical Statistics *21*, 309.

[95] Henderson, C., Kempthorne, O., Searle, S., and von Krosigk, C. M. (1959). The estimation of environmental and genetic trends from records subject to culling. Biometrics *15*, 192–218.

[96] Henderson, C. (1963). Selection index and expected genetic advance. Statistical Genetics and Plant Breeding pp. 141–163.

[97] Hastie, T., Tibshirani, R., and Friedman, J. (2001). The elements of statistical learning - data mining, inference and prediction. (Springer New York).

[98] Abney, M., McPeek, M. S., and Ober, C. (2000). Estimation of variance components of quantitative traits in inbred populations. American journal of human genetics *66*, 629–50.

[99] Ober, C., Abney, M., and McPeek, M. (2001). The genetic dissection of

complex traits in a founder population. The American Journal of Human Genetics *69*, 1068–1079.

[100] Loiselle, B., Sork, V., Nason, J., and Graham, C. (1995). Spatial genetic structure of a tropical understory shrub, Psychotria officinalis (Rubiaceae). American Journal of Botany *82*, 1420–1425.

[101] Hardy, O. and Vekemans, X. (2002). SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Molecular Ecology Notes *2*, 618–620.

[102] Balding, D. and Nichols, R. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica *96*, 3–12.

[103] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression: Rejoinder. Annals of Statistics *32*, 494–499.

[104] Club, A. K. (1998). The complete dog book. (Howell Book House).

[105] Darwin, C. (1859). On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. New York: D. Appleton.

[106] Wayne, R. (1986). Cranial morphology of domestic and wild canids: the influence of development on morphological change. Evolution *40*, 243–261.

[107] Wayne, R. (1986). Limb morphology of domestic and wild canids: the influence of development on morphologic change. Journal of Morphology *187*, 301–319.

[108] Sutter, N., Mosher, D., Gray, M., and Ostrander, E. (2008). Morphometrics within dog breeds are highly reproducible and dispute rensch's rule. Mammalian Genome *19*, 713–723.

[109] Candille, S., Kaelin, C., Cattanach, B., Yu, B., Thompson, D., Nix, M., Kerns, J., Schmutz, S., Millhauser, G., and Barsh, G. (2007). A beta-defensin mutation causes black coat color in domestic dogs. Science *318*, 1418.

[110] Karlsson, E., Baranowska, I., Wade, C., Hillbertz, N., Zody, M., Anderson,

N., Biagi, T., Patterson, N., Pielberg, G., Kulbokas, E., et al. (2007). Efficient mapping of mendelian traits in dogs through genome-wide association. Nature genetics *39*, 1321–1328.

[111] Parker, H., VonHoldt, B., Quignon, P., Margulies, E., Shao, S., Mosher, D., Spady, T., Elkahloun, A., Cargill, M., Jones, P., et al. (2009). An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs. Science *325*, 995.

[112] Drogemuller, C., Karlsson, E., Hytonen, M., Perloski, M., Dolf, G., Sainio, K., Lohi, H., Lindblad-Toh, K., and Leeb, T. (2008). A mutation in hairless dogs implicates foxi3 in ectodermal development. Science *321*, 1462.

[113] Akey, J., Ruhe, A., Akey, D., Wong, A., Connelly, C., Madeoy, J., Nicholas, T., and Neff, M. (2010). Tracking footprints of artificial selection in the dog genome. Proceedings of the National Academy of Sciences *107*, 1160.

[114] Cadieu, E., Neff, M. W., Quignon, P., Walsh, K., Chase, K., Parker, H. G., Vonholdt, B. M., Rhue, A., Boyko, A., Byers, A., et al. (2009). Coat variation in the domestic dog is governed by variants in three genes. Science *326*, 150–3.

[115] Hillbertz, N., Isaksson, M., Karlsson, E., Hellmén, E., Pielberg, G., Savolainen, P., Wade, C., von Euler, H., Gustafson, U., Hedhammar, A., et al. (2007). Duplication of FGF3, FGF4, FGF19 and ORAOV1 causes hair ridge and predisposition to dermoid sinus in Ridgeback dogs. Nature genetics *39*, 1318–1320.

[116] Weedon, M. N., Lango, H., Lindgren, C. M., Wallace, C., Evans, D. M., Mangino, M., Freathy, R. M., Perry, J. R. B., Stevens, S., Hall, A. S., et al. (2008). Genome-wide association analysis identifies 20 loci that influence adult height. Nature genetics *40*, 575–83.

[117] Boyko, A. R., Boyko, R. H., Boyko, C. M., Parker, H. G., Castelhano, M., Corey, L., Degenhardt, J. D., Auton, A., Hedimbi, M., Kityo, R., et al. (2009). Complex population structure in african village dogs and its implications for inferring dog domestication history. Proc Natl Acad Sci USA *106*, 13903–8.

[118] Quignon, P. (2010). Personal communications.

[119] Chase, K., Jones, P., Martin, A., Ostrander, E. A., and Lark, K. G. (2009).

Genetic mapping of fixed phenotypes: Disease frequency as a breed characteristic. Journal of Heredity *100*, S37–S41.